

Comparison of various statistical methods for identifying differential gene expression in replicated microarray data

Seo Young Kim Research Institute for Basic Science, Chonnam National University, Gwangju, Korea and **Jae Won Lee, In Suk Sohn** Department of Statistics, Korea University, Seoul, Korea

DNA microarray is a new tool in biotechnology, which allows the simultaneous monitoring of thousands of gene expression in cells. The goal of differential gene expression analysis is to identify those genes whose expression levels change significantly by the experimental conditions. Although various statistical methods have been suggested to confirm differential gene expression, only a few studies compared the performance of the statistical tests. In our study, we extensively compared three types of parametric methods such as T-test, B-statistic and Bayes T-test and three types of non-parametric methods such as samroc, significance analysis of microarray and a modified mixture model using both the simulated datasets and the three real microarray experiments.

1 Introduction

DNA microarray is a new tool in biotechnology, which allows the simultaneous monitoring of thousands of gene expression in cells.¹ One important and common question in microarray experiments is how the identification of differential gene expression can be made. A common task in analyzing microarray data is to determine which genes are differentially expressed across two tissue samples or samples obtained under two experimental conditions. Microarrays present new statistical problems because the data are very highly dimensionalized with a very small number of replications. Recently, several statistical methods have been proposed to handle this situation when there are replicated microarray data collected from various experimental conditions.

The fold change rule, which relies on the fold increase/decrease cutoff to identify differentially expressed genes, has been widely used in analyzing the microarray data. It is also defined in Section 2.7. For example, Schena *et al.*² declared a gene differentially expressed if its expression level differs by more than 5-fold in the two mRNA samples. DeRisi *et al.*³ identified differentially expressed genes using a 3-fold for the log ratios of the expression levels. A basic statistical problem is to know when the measured differential expression is likely to reflect a real biological shift in gene expression. This

Address for correspondence: Jae Won Lee, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jael@korea.ac.kr

depends on the amount of variations in the system, and thus it is difficult to justify a fixed rule. It has been known that simply using the arbitrary fold change such as 3- and 5-fold changes is unreliable and inefficient.⁴

Many statistical approaches have been proposed to model some distributional properties of gene expression.^{2–13} Their approaches can be divided into two groups according to the number of arrays handled. The first group analyzes the single microarray containing only one spot per each gene, which needs very strong parametric assumptions, and the second group handles multiple microarrays. Lee *et al.*¹⁴ reported that the single microarray data tend to include high noises that make the result unreliable. Kerr and Churchill¹⁵ also stressed the importance of replications in the microarray analysis, and it contributed to develop various statistical methods for dealing with multiple microarray data.

These statistical methods can be divided into parametric and non-parametric approaches. Long *et al.*¹⁶ applied the traditional T-test based on a Bayesian estimate of variance among the experiment replicates with normally distributed expression measurements, and Baldi and Long¹⁷ used a hierarchical Bayesian modeling framework with normal models. Newton *et al.*⁷ proposed the Bayesian model based approach in the analysis of differential expression, and Lönnstedt and Speed¹⁰ used a parametric empirical Bayes method to analyze the replicated microarray data. Smyth¹¹ developed the hierarchical model of Lönnstedt and Speed into a practical approach. All these approaches used the arbitrary cutoff points and/or probabilistic inferences based on the specific model. However, the strong normality assumption could be violated in many real situations, and it is important to check the robustness of these methods. The non-parametric approaches have been also used. Efron *et al.*⁵ adopted an empirical Bayes method, Tusher *et al.*,⁸ utilized the significance analysis of microarray (SAM) method, Dudoit *et al.*¹² reported the non-parametric T-test with adjusted P-value, and Pan *et al.*¹⁸ also reported the use of mixture modeling method (MMM). Recently, Pan⁹ compared the MMM with two other parametric methods such as the traditional T-test and the regression modeling method by Thomas *et al.*¹⁹ The basic idea of these non-parametric approaches is to estimate null distribution of the test statistic directly rather than assuming the null hypothesis of no differential gene expression. But the non-parametric SAM method by Tusher *et al.*,⁸ empirical Bayes method by Efron *et al.*⁵ and MMM by Pan *et al.*¹⁸ still assumed that the denominator and the numerator of the null statistic are independent. This assumption can be violated for some null statistics. To overcome this problem, Zhao and Pan¹³ proposed a modified MMM, and Broberg²⁰ ranked genes in the order of likelihood of being differentially expressed.

Although various statistical methods have been suggested to confirm the differential gene expression, only a few studies have compared the performances of the different statistical approaches. Thus, it is very important to distinguish their merits and demerits through extensive and fair comparisons. A comparative study of the non-parametric methods by Troyanskaya *et al.*²¹ a review study of the parametric methods by Smyth *et al.*²² and a comparative study by Broberg²⁰ have been recently conducted. Broberg²⁰ has proposed the samroc method and compared it with several methods such

as SAM by Tusher *et al.*⁸ B-statistic by Smyth,¹¹ T-test and fold change rule. As these statistical approaches have been widely used, we need to establish more balanced point of view through more extensive comparative study.

In our study, we compare three types of parametric methods such as T-test, B-statistic and Bayes T-test and three types of non-parametric methods such as samroc, SAM, and a modified mixture model proposed by Zhao and Pan.¹³ In Section 2, we review in detail these different statistical methods for identifying differentially expressed genes. In Section 3, we describe three datasets along with preliminary data processing procedure and present the comparison results in Section 4. Finally, we conclude with an extensive discussion in Section 5.

2 Statistical methods for identifying differential gene expression

In this section, we review several statistical methods for determining differential expression in microarray data. Suppose that the experimental data consist of measurements y_{gi} under two conditions, where i ($=1, 2, \dots, k$) denotes the i -th array, g ($=1, 2, \dots, G$) denotes the g -th gene, and k_1 and k_2 are the number of arrays for each condition, that is, $k = k_1 + k_2$. Let the sample means and the sample variances of y_{gi} 's for gene g under two conditions be denoted as \bar{y}_{g1}, s_{g1}^2 and \bar{y}_{g2}, s_{g2}^2 , respectively. Here, diff is the difference between \bar{y}_{g1} and \bar{y}_{g2} , and s_g and Se_g represent the pooled standard deviation and the standard error of the diff across the replicates for the gene, respectively.

2.1 T-statistic

The two sample T-statistic with two independent normal samples without assuming the equal variances between two samples could be written as follows;

$$t_g = \frac{\text{diff}}{Se_g}, \quad Se_g = \sqrt{\frac{s_{g1}^2}{k_1} + \frac{s_{g2}^2}{k_2}}$$

A gene with very small variance due to its low expression level contributes to have large absolute t -value regardless of the mean difference under two conditions, and thus this gene can be selected as the differentially expressed genes although they are not truly differentially expressed. To overcome this problem of the traditional T-test, various methods have been proposed. Among these methods, there are SAM proposed by Tusher *et al.*, B-statistic proposed by Smyth and samroc proposed by Broberg.^{8,11,20}

2.2 B-statistic

Lönnstedt and Speed¹⁰ proposed a statistic based on the empirical Bayes approach which employs a Bayes log posterior odds. It avoids the problems of traditional T-statistic through the log odds posterior statistic under the conditional normality assumption. Smyth¹¹ developed the hierarchical model of Lönnstedt and Speed into a practical approach and proposed the B-statistic. The model is reset in the context of general

linear model with arbitrary coefficients and contrast of interest. Smyth's B-statistic is defined as

$$B = \frac{p(\beta_g \neq 0 | \tilde{t}_g, s_g^2)}{p(\beta_g = 0 | \tilde{t}_g, s_g^2)} = \frac{p}{1-p} \left\{ \frac{\nu_g}{\nu_g + \nu_0} \right\}^{1/2} \left\{ \frac{\tilde{t}_g^2 + d_0 + d_g}{\tilde{t}_g^2 (\nu_g / (\nu_g + \nu_0)) + d_0 + d_g} \right\}^{(1+d_0+d_g)/2}$$

where β_g is the regression coefficient of the contrast of interest, that is, control versus experimental group, and the contrast estimator $\hat{\beta}_g$ is assumed to be approximately normal with mean β_g and covariance matrix $C^T V_g C \sigma_g^2$. Here, V_g is called unscaled covariance matrix, s_g^2 is residual sample standard deviation as an estimator of σ_g^2 , ν_g is the diagonal element of $C^T V_g C$, and d_g is the residual degree of freedom for the linear model for gene g . ν_0 and d_0 are prior estimators of ν_g and d_g , respectively, and they can be estimated from the data. p is the expected proportion of truly differentially expressed genes.

To compute the previous posterior odds, Smyth proposed a 'moderated' T- statistic

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{\nu_g}}, \quad \tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

where $\hat{\beta}_g$ is a regression coefficient estimate for the contrast of interest. Here, \tilde{s}_g^{-2} is the posterior mean of σ_g^{-2} given s_g^2 under the hierarchical model, and s_0^2 with degree of freedom d_0 is a prior estimator of s_g^2 and also can be estimated from s_g^2 . Smyth also stated that the \tilde{t}_g shows robust behavior for small sample case and is designed to accommodate different variances. On the other hand, \tilde{t}_g depends on d_0 and s_0^2 . Provided that $d_0 < \infty$ and $d_g > 0$, \tilde{t}_g is similar to traditional T-statistic. B-statistic still depends on hyperparameters ν_0 , d_0 , s_0^2 , and p , and these parameters can be estimated from the data.

2.3 Bayes T-test

Baldi and Long¹⁷ developed a Bayesian probabilistic framework for microarray data analysis. Their statistic is used to solve small variance problems in low expression level and uses the parametric Bayesian method to have the parameters (mean, standard deviation and so on.) for T-statistic. This statistic is well known for its effectiveness in analyzing the samples having small size, but it still heavily depends on the parametric assumption. Bayes T-test uses the estimate of parameters such as population mean (μ) and variance (σ^2) by Bayesian method instead of sample mean and sample variance of the traditional T-statistic. The mean of posterior estimate in each group is given as

$$\mu_j = \mu_{nj}, \quad \sigma_j^2 = \frac{\nu_j \sigma_{nj}^2}{\nu_j - 2}$$

where the mean of the posterior estimate (μ_{nj}) is a convex weighted average of the prior mean (μ_{0j}) and the sample mean (\bar{y}_j) for group j , $j = 1, 2$, that is,

$$\mu_{nj} = \frac{\lambda_{0j}}{\lambda_{0j} + k_j} \mu_{0j} + \frac{k_j}{\lambda_{0j} + k_j} \bar{y}_j$$

The hyperparameters μ_{0j} and σ_j^2/λ_{0j} can be interpreted as the location and the scale of μ_j , respectively, and k_j is the sample size for each group. σ_{nj}^2 is posterior variance component and posterior sum of squares is $v_j \sigma_{nj}^2 = v_{0j} \sigma_{0j}^2 + (k_j - 1) s_j^2 + \lambda_{0j} k_j / (\lambda_{0j} + k_j) (\bar{y}_j - \mu_{0j})^2$, and the posterior degree of freedom is $v_j = v_{0j} + k_j$. In Bayes T-test, the hyperparameters for the prior v_{0j} and σ_{0j}^2 can be interpreted as the degree of freedom and scale of σ_j^2 , respectively. Owing to the complicated theoretical background, we will not discuss it here in more detail. This statistic is currently implemented in the Cyber-T software, which is accessible at www.genomics.uci.edu/software.html.

2.4 Significant analysis of microarrays (SAM)

To avoid the small variance problem of T-test, SAM uses a statistic similar to T-statistic and the permutation of repeated measurements to estimate the false discovery rate.⁸ At low expression levels, the absolute value of t_g can be high because of small values in Se_g . The shortcoming of the traditional T-test is that genes with small sample variances due to the low expression levels have high chance of being declared as the differentially expressed genes. Thus SAM added a small positive constant a to alleviate this problem. The SAM statistic is

$$t_{\text{sam}} = \frac{\text{diff}}{Se_g + a}, \quad Se_g = s_g \sqrt{\frac{1}{k_1} + \frac{1}{k_2}}$$

where the value for a is chosen to minimize the coefficient of variation. SAM is similar to the method by Efron *et al.*,⁵ which use a to be equal to the 90th percentile of the standard errors of all the genes. SAM assigns a score based on changes that is related to the standard deviation of repeated measurements for that gene. Genes with scores greater than a cutoff value are determined to be significant.

2.5 Samroc

Broberg²⁰ proposed a method for ranking genes in the order of likelihood of being differentially expressed, which is often called as samroc. The main purpose of this method is to estimate the false negative (FN) and false positive (FP) rates. The procedure sets out to minimize these errors. The samroc method is similar to SAM, although an added constant in the denominator of the statistic is different. The proposed statistic is

$$t_{\text{samroc}} = \frac{\text{diff}}{Se_g + b}$$

Main interest is to find the optimal constant b for given significance level of α . This procedure proposed a criterion, which is the distance of points on the curve to the origin, for choosing a good receiver operating characteristic (ROC) curve. ROC curve allows users to compare the FP error rate and FN error rate of various test statistics without involving P -values. This minimizes the number of genes that are falsely declared positive and falsely declared negative for a given significance level of α and a value b .

2.6 Zhao-Pan method

Zhao and Pan¹³ adopted a modified non-parametric approach to detect the differentially expressed genes in replicated microarray experiments. The basic idea of this non-parametric method lies in estimating the null distribution of test statistic, say Z_g , by directly constructing a null statistic, say z_g , such that the distribution of z_g is the same as the distribution of Z_g under the null hypothesis. This avoids the strong assumptions about the null distribution of the parametric methods. Among the earlier studies, there are empirical Bayes method by Efron *et al.*,⁵ SAM by Tusher *et al.*,⁸ and a MMM by Pan *et al.*¹⁸ However, a common problem with these methods is that the numerator and the denominator of z_g and Z_g are assumed to be independent of each other. In practice, this independency is violated by z_g , and Zhao and Pan¹³ use z_g and Z_g as below to overcome this problem. At $k_1 > 2k_2, i = 1, 2, \dots, k_1, k_1 + 1, \dots, k_1 + k_2$,

$$Z_g = \frac{\bar{y}_{g(1a)} - \bar{y}_{g(2)}}{\sqrt{s_{g(1a)}^2/(k_1 - k_2) + s_{g(2)}^2/k_2}}, \quad z_g = \frac{\bar{y}_{g(1a)} - \bar{y}_{g(1b)}}{\sqrt{s_{g(1a)}^2/(k_1 - k_2) + s_{g(1b)}^2/k_2}}$$

where

$$\bar{y}_{g(1a)} = \frac{\sum_{i=1}^{k_1-k_2} y_{gi}}{k_1 - k_2}, \quad s_{g(1a)}^2 = \frac{\sum_{i=1}^{k_1-k_2} (y_{gi} - \bar{y}_{g(1a)})^2}{k_1 - k_2 - 1}$$

$$\bar{y}_{g(1b)} = \frac{\sum_{i=k_1-k_2+1}^{k_1} y_{gi}}{k_2}, \quad s_{g(1b)}^2 = \frac{\sum_{i=k_1-k_2+1}^{k_1} (y_{gi} - \bar{y}_{g(1b)})^2}{k_2 - 1}$$

and

$$\bar{y}_{g(2)} = \frac{\sum_{i=k_1+1}^{k_1+k_2} y_{gi}}{k_2}, \quad s_{g(2)}^2 = \frac{\sum_{i=k_1+1}^{k_1+k_2} (y_{gi} - \bar{y}_{g(2)})^2}{k_2 - 1}$$

For the case of $k_1 \geq k_2$, but $k_1 < 2k_2$, the reader can refer to the Zhao and Pan.¹³

2.7 Fold change rule

Fold change is defined as

$$FC = \frac{\max(\bar{y}_{g1}, \bar{y}_{g2})}{\min(\bar{y}_{g1}, \bar{y}_{g2})}$$

Table 1 Properties of the six testing methods

Method	Sample size	Distributional assumption	Equal variance assumptions between groups
SAM	Small	None	Equal
Samroc	Small	None	Equal
B-statistic	Small	Strong	Unequal
T-statistic	Large	Strong	Unequal
Zhao–Pan	Large	Weak	Equal
Bayes T-test	Small	Strong	Unequal

where \bar{y}_g is the mean of the intensity measurements for the g -th gene. According to the m -fold change rule, the g -th gene is declared as significantly differentially expressed gene if $FC > m$ or $FC < 1/m$.

Table 1 summarizes main features of the previous methods in the context of sample size, distributional assumption, and variance condition between two groups. In general, SAM, samroc, B-statistic and Bayes T-test are known to work well with the small sample size, and T-statistic and Zhao–Pan method are known to perform well with large sample size. This difference may be related to the fact that SAM and samroc do not need any distributional assumption, whereas the others need distributional assumptions for the analysis. Of these six methods, SAM, samroc and Zhao–Pan method require the equal variance assumption between two groups.

3 Real datasets

3.1 Leukemia data

We used the leukemia dataset of Golub *et al.*²³, which consists of 38 learning samples on the Affymetrix high density oligonucleotide chips containing $G = 7129$ human genes. The goal of this experiment is to identify genes that are differentially expressed in 27 acute lymphoblastic leukemia (ALL) patients and 11 acute myeloid leukemia (AML) patients. Of 38 samples, we selected 26 ALL samples and 10 AML samples as in Broberg.²⁰ The data were preprocessed by subtracting the median and dividing its interquartile range ($IQR = \text{upper quartile} - \text{lower quartile}$) as in Broberg.²⁰ This dataset is available at <http://www.genome.wi.mit.edu/MPR>.

3.2 Melanoma data

The melanoma dataset is described in Bittner *et al.*,²⁴ and available at http://www.nhgri.nih.gov/DIR/Microarray/Melanoma_Supplement/index.html. This dataset was acquired from a study of gene expression in two types of 31 cutaneous melanomas and 7 controls. Gene expression levels were measured using cDNA microarrays containing $G = 8150$ human genes. In each microarray experiment, fluorescent cDNA targets were prepared from an experimental mRNA sample (Cy5), and a reference mRNA sample derived from a single probe labeled as Cy3 was used for all 38 samples. Of the 8150 genes, 3613 genes were identified as well measured. This experiment data had many Cy5/Cy3 expression ratios above 10 000 and also had many below 0.02. Therefore, the

data filtering method, which excluded the genes with expression ratio greater than 50 and less than 0.02, was applied for this dataset as in Darlene *et al.*²⁵ These ratios were transformed to a base 2 logarithmic scales, and the data were also preprocessed by subtracting the median and dividing its IQR as described in Broberg.²⁰

3.3 Apo AI mouse data

We applied the earlier methods to the apolipoprotein AI (Apo AI) dataset of Callow *et al.*,²⁶ which consists of treatment group of eight mice with the Apo AI gene knocked out and control group of eight normal mice. The Apo AI knockout is a mouse model with extremely low high density lipoprotein (HDL) cholesterol levels, but this protein acts via different mechanisms to affect HDL cholesterol delivery to the liver. The goal of this experiment is to find genes that were differentially expressed in the livers of treatment mice compared to control mice. For these 16 mice, target sample used all red fluorescent dye (Cy5), and reference sample used all green dye (Cy3). We set $k_1 = 8, k_2 = 8$. There are $G = 6384$ genes in each sample that were preprocessed by subtracting the median and dividing its IQR as in Broberg.¹⁸ For each 16 microarrays, we evaluated the log intensity ratio, $\log_2 \text{Cy5/Cy3}$, for gene g . In this experiment, eight significant genes which correspond to only four distinct genes, that is, Apo AI (three copies), Apo CIII (two copies), sterol C5 (two copies) and a novel EST, were found. All the changes were confirmed by RT-PCR in Callow *et al.*²⁶

4 Simulation results

4.1 Simulated data results

We carried out an extensive simulation study to evaluate each of the earlier methods. Lönnstedt and Speed, Baldi and Long, and Storey^{10,17,27} have conducted other simulation studies. In this article, we simulated artificial dataset assuming normal and lognormal distributions of log expression levels as in Baldi and Long.¹⁷ Baldi and Long generated a simulated dataset assuming normal distribution of log expression, with means and variances similar to those in the *E. coli* cDNA data of Arfin *et al.*²⁸ The means and the standard deviations for the simulation data are shown in Table 2. In Table 2, the first three rows represent the unchanged genes, whereas the last three rows represent the changed genes. We generated 10 000 genes from selected large sample size (26 & 10 arrays) and small samples (4 & 4 arrays) as in the real leukemia data. Because SAM, samroc, B-statistic and the Zhao–Pan method require equal variance under the null hypothesis, to check their robustness to the assumption violation, we also considered the case where the two distributions are normally distributed with different variances as shown in Table 3. In both tables, the values of means and standard deviations were chosen to control the coefficient of variation between two groups.

The sample sizes were (26 & 10) and (4 & 4) arrays, which are the same as those for the real leukemia dataset. From these results, we compared the number of true positive genes and the average ranks for various methods among the top 500 ranked genes. The selection of 500 ranked genes is of course arbitrary, but resembles real biological situation

Table 2 The means and standard deviations for the simulated data under normal distributions

Mean 1	Sd 1	Mean 2	Sd 2
-8	0.4	-8	0.4
-10	0.8	-10	0.8
-12	1.0	-12	1.0
-6	0.2	-6.1	0.2
-8	0.4	-8.5	0.5
-10	0.8	-11	1.0

Table 3 The means and standard deviations for the simulated data under normal distributions when the variances are different under the null hypothesis

Mean 1	Sd 1	Mean 2	Sd 2
-8	0.4	-8	0.4
-10	0.8	-10	1.0
-12	1.0	-12	1.5
-6	0.2	-6.4	0.2
-8	0.4	-9	0.6
-10	0.8	-12	1.2

as in Broberg.²⁰ Simulated data contained 5% changed genes out of 10 000 genes, and Tables 4 and 5 show the main results of simulation study.

4.1.1 Dataset with almost equal variance between two groups

Table 4 shows the simulation results when the two groups have almost equal variance given in Table 2. When the data are truly normal and there are 26 and 10 samples in each group, SAM, samroc and B-statistic perform well. For lognormal data, T-statistic, SAM together with samroc perform well. For the dataset containing 4 arrays per each group, Bayes T-test appears to be the best in both normal and lognormal cases. Also, samroc

Table 4 The number of true positives and the average ranks among the top 500 when the proportion changed equals 5% and the number of genes equals 10 000 (true positives: average ranks)

Method	(26,10) arrays		(4,4) arrays	
	Normal	Lognormal	Normal	Lognormal
B-statistic	484:255.6	425:340.4	220:1264.1	277:1041.4
SAM	484:254.6	455:287.8	227:1124.5	279:860.7
Samroc	489:254.1	461:288.4	288:768.6	280:989.6
T-statistic	467:265.1	460:307.1	269:983.1	259:1170.2
Zhao-Pan	462:270.4	443:322.4	104:2308.0	103:2379.3
Bayes T-test	462:285.0	409:348.4	351:849.6	363:883.6
Fold change	434:286.3	278:827.9	196:1257.2	107:2417.1

Table 5 The number of true positives and the average ranks among the top 500 when the proportion changed equals 5%, the number of genes equals 10 000 and the variances are different under the null hypothesis (true positives: average ranks)

Method	(26,10) arrays		(4,4) arrays	
	Normal	Lognormal	Normal	Lognormal
B-statistic	447:287.69	338:551.46	167:1777.73	228:1545.65
SAM	438:294.70	342:532.43	185:1278.40	224:1500.62
Samroc	453:285.48	336:569.50	255:1130.20	233:1400.86
T-statistic	433:317.96	419:425.35	233:1393.38	217:1686.84
Zhao–Pan	420:348.70	391:500.19	79:2841.15	78:2969.57
Bayes T-test	467:270.53	214:1680.79	271:1179.02	278:1567.17
Fold change	190:1535.90	156:1766.18	162:2058.91	86:3285.33

and T-statistic perform well in the normal data case, and SAM, samroc and B-statistic perform quite well in the lognormal case.

4.1.2 Dataset with unequal variances between two groups

Table 5 shows the simulation results when the two groups have unequal variances given in Table 3. As shown in Table 5, for the (26 & 10) arrays, Bayes T-test appears to be the best and samroc, B-statistic and SAM also perform well in the normal data case, whereas T-statistic and the Zhao–Pan statistic perform quite well in lognormal data case. For the (4 & 4) arrays, Bayes T-test and samroc have good performances in both normal and lognormal data cases, although the performance of SAM and B-statistic are also good in lognormal data case.

Broberg²⁰ suggested samroc as the best method through simulation study which included only the small sample case. We extended Broberg’s simulation study to allow for both large and small sample cases and also for both equal and unequal variances between two groups, and compared six testing methods on a fair basis. It is also interesting to see how each testing method responds to varying variances and sample sizes. Zhao and Pan proposed that Zhao–Pan test is more preferable with small sample size,¹³ but the performance of Zhao–Pan test seems to be poor in the small sample case. Both SAM and samroc perform well in most cases, but samroc seems to be slightly better than SAM when the sample size is small. However, SAM gives the slightly better result than samroc when the sample size is small, distribution is lognormal and the variances are unequal.

4.2 Real microarray data results

The statistical methods described in Section 2 were applied to the three real microarray datasets discussed in Section 3. To compare six testing methods, we used the average ranks of the reference genes which are believed to be truly differentially expressed, and thus the choices of reference genes are quite critical in the comparison study.

Broberg²⁰ adopted 50 reference genes that were selected by the MMM of Pan *et al.*¹⁸ in the leukemia data, ranked all the genes in the order of large absolute values of each test statistic, and compared the average ranks of the four testing methods such as B-statistic, SAM, samroc and T-statistic. The adopted 50 reference genes which were used in Broberg

consist of highly expressed top 25 genes and bottom 25 genes in AML compared to ALL.²⁰ Thus, the comparison of average rankings of these 50 reference genes would be essentially the same as comparison of the test results to that of MMM. Broberg used the leukemia dataset and reported samroc to be the best method in small sample case. However, we can interpret that samroc will deliver the most similar test results as MMM, and the use of 50 reference genes that were selected from MMM seems to be unfair to compare the performances of the testing methods based on the average ranks. Therefore, in our study, we adopted a different approach in contrast to Broberg’s reference gene selection. We used the reference genes which show significant difference between two samples by all the tests such as B-statistic, SAM, samroc, T-statistic, Zhao–Pan and Bayes T-test methods. We initially selected top 5% significant genes by each of six testing methods and finally selected a small number of reference genes (60 in leukemia, 69 in melanoma and 27 in mouse dataset) that were commonly found to be significant by all the six methods. Table 6 shows the average ranks of the reference genes in both large and small sample cases.

For the leukemia dataset, we used both large sample (26 & 10 arrays) and small sample (4 & 4 arrays) per group. As mentioned earlier, we initially selected 356 significant genes (i.e., 5% of a total of 7129 genes) from each method, and finally selected 60 reference genes that were commonly found to be significant by all the six methods. As shown in Table 6, B-statistic and SAM give the lowest and the second lowest average rank in both large and small sample cases, respectively. Note that Broberg also used this small sample leukemia data to compare B-statistic, SAM, samroc and T-statistic and claimed that samroc gives the lowest average rank.²⁰ There is no doubt that this difference comes from the different selection of the reference genes.

For the melanoma dataset, we used both large sample (31 & 7 arrays) and small sample (4 & 4 arrays) per group. As mentioned earlier, we initially selected 407 significant genes (i.e., 5% of a total of 8150 genes) from each method, and finally selected 69 reference genes that were commonly found to be significant by all the six methods. In large sample case, B-statistic, SAM, samroc and T-statistic give almost the same average ranks and perform better than Zhao–Pan method and Bayes T-test. In small sample case, however, SAM performs much better than the other tests, and B-statistic and T-statistic also perform better than samroc, Bayes T-test or Zhao–Pan method.

For Apo AI mouse dataset, we used both large sample (8 & 8 arrays) and small sample (4 & 4 arrays) per group. As mentioned earlier, we initially selected 319 significant genes (i.e., 5% of a total of 6384 genes) from each method, and finally selected 27 reference

Table 6 Average ranks of the reference genes in each dataset

		SAM	B-statistic	Samroc	T-statistic	Zhao–Pan	Bayes T-test
Leukemia	Large (26 & 10)	69.08	66.2	85.03	134.46	128.48	129.96
	Small (4 & 4)	482.08	439.7	562.63	532.58	2289.85	701.65
Melanoma	Large (31 & 7)	119.01	116.01	120.13	120.36	133.07	237.21
	Small (4 & 4)	772.0	1119.28	1318.02	1177.24	1446.88	1382.63
Apo AI	Large (8 & 8)	65.44	75.88	57.03	58.70	98.66	87.03
	Small (4 & 4)	725.11	738.62	712.81	711.37	929.66	683.44

genes that were commonly found to be significant by all the six methods. In large sample case, samroc and T-statistic perform the best, and SAM and B-statistic also perform well. In small sample case, however, Bayes T-test performs the best, and samroc and T-statistic also perform well.

Thus far, we have compared the performances of six different methods by using the reference genes from each dataset. We have seen that SAM and B-statistic give consistently good performance regardless of the sample size in the leukemia and melanoma datasets, whereas samroc and T-statistic perform the best in Apo AI mouse dataset. Although Broberg reported samroc as the best testing method among SAM, B-statistic, samroc and T-statistic,²⁰ samroc gives the satisfactory result only in Apo AI mouse data in our study. Therefore, our analysis of three different microarray datasets using the different reference gene sets show that the performance of the testing procedure also depends on the dataset and thus Broberg's result using the leukemia dataset only cannot be easily generalized to the other types of microarray data.

5 Conclusion and Discussion

We have compared the performances of six well-known testing procedures such as T-statistic, SAM, samroc, B-statistic, Zhao–Pan and Bayes T-test for identifying the differentially expressed genes using the microarray datasets. All the testing methods were compared using both the real gene expression datasets and the simulated datasets.

Through the analysis of three real datasets, we are able to recognize that each testing method gives different results depending upon the microarray data. One primary reason could be explained by the fact that each testing method selects different top ranking genes from the same dataset. In addition, the performance of the testing methods depends on the normal distribution assumption or equal variance assumption of the log ratios of expression levels. For each of the three real datasets, we checked the normality assumption by Kolmogorov-Smirnov test and the equal variance assumption by F-test. Table 7

Table 7 Number of genes satisfying normality assumption and/or equal variances assumption at each dataset

Dataset	Assumptions	Number of genes satisfying assumption (%)	Number of genes which do not satisfy assumption (%)	Total (%)
Leukemia	Normality	2246 (31.5)	4883 (68.5)	7129 (100)
	Equal Variance	1690 (23.7)	5439 (76.3)	7129 (100)
Melanoma	Normality	1311 (36.3)	2302 (63.7)	3613 (100)
	Equal Variance	874 (24.2)	2739 (75.8)	3613 (100)
Apo AI	Normality	5011 (78.5)	1373 (21.5)	6384 (100)
	Equal Variance	5426 (85.0)	958 (15.0)	6384 (100)

shows the number of genes which follow these assumptions. In the majority of genes, the leukemia data follow non-normal distribution (68.5%) and have unequal variances (76.3%), the melanoma data follow non-normal distribution (63.7%) and have equal variance (75.8%), and the Apo AI data follow normal distribution (78.5%) and have equal variance (85%). For illustrative purpose, we also selected two genes from each of the three real datasets and drew the quantile–quantile (Q–Q) plots of the log ratios of expression levels to check the normality assumption (see Figure 1). In Figure 1, plot I–II, III–IV and V–VI are from the leukemia, melanoma, and Apo AI data, respectively.

As discussed earlier, Broberg proposed *samroc* to be the best testing method among SAM, *samroc*, B-statistic and T-statistic.²⁰ The basic idea of Broberg's study is to use MMM method in selecting 50 reference genes for the analysis of real data and to use the average rank for comparing the performance of the testing methods. We can point out two problems in Broberg's study. First, this study practically failed to select fair reference genes. In other words, the use of Pan's MMM to select reference genes for comparing six testing methods gives the best performance of the testing method which is most similar to Pan's MMM method. Secondly, the adopted 50 reference genes which were used in Broberg consist of highly expressed top 25 genes and bottom 25 genes in AML compared to ALL.²⁰ That is, they are not actually the top 50 genes that were selected in the order of large absolute value of the MMM test statistic. Ranking all the genes in the order of large absolute value change the ranks of the reference genes and their average rank. To see the effect of the gene selection, we also selected 50 reference genes in the order of large absolute values of the MMM test statistics instead of 25 positive and 25 negative values, and compared the average ranks of the testing methods when only the first four samples from ALL and AML were used. Table 8 shows the ranks of 50 reference genes and their average ranks when each of the six testing methods was applied. Among the 50 reference genes that were selected in Table 7, 30 genes were also found in Broberg's set of 50 reference genes. We found that all the 25 positive-valued genes in Broberg's set of 50 reference genes were also included in our reference gene set, but only five negative-valued genes were included in our 50 reference gene set. Table 8 shows that the B-statistic, SAM and *samroc* all performed well although B-statistic and SAM are slightly better than *samroc*.

Small sample size and the skewed data are common in the microarray experiments because the microarray data represent various biological situations, and thus it is important to see how the performance of the testing methods can be affected by those factors. Through our comparison study, we can find that the sample size, distributional assumption and the assumption of equal variance between two groups under the null hypothesis make non-ignorable effects on the performance of the testing methods. Although SAM, *samroc*, B-statistic and Zhao–Pan method require equal variance under the null hypothesis, SAM, *samroc* and B-statistic do not seem to be sensitive to the violation of equal variance assumption. SAM, *samroc* and B-statistic perform well and Zhao–Pan method performs poorly in most cases. Especially, Zhao–Pan method is strongly discouraged to use in the small sample case. In both large and small sample cases, *samroc* seems to perform slightly better than SAM or B-statistic in finding the significant genes in the simulation study. In the real data analysis, *samroc* also gives the lower average ranks than SAM or B-statistic in Apo AI mouse data which follows the normal distribution,

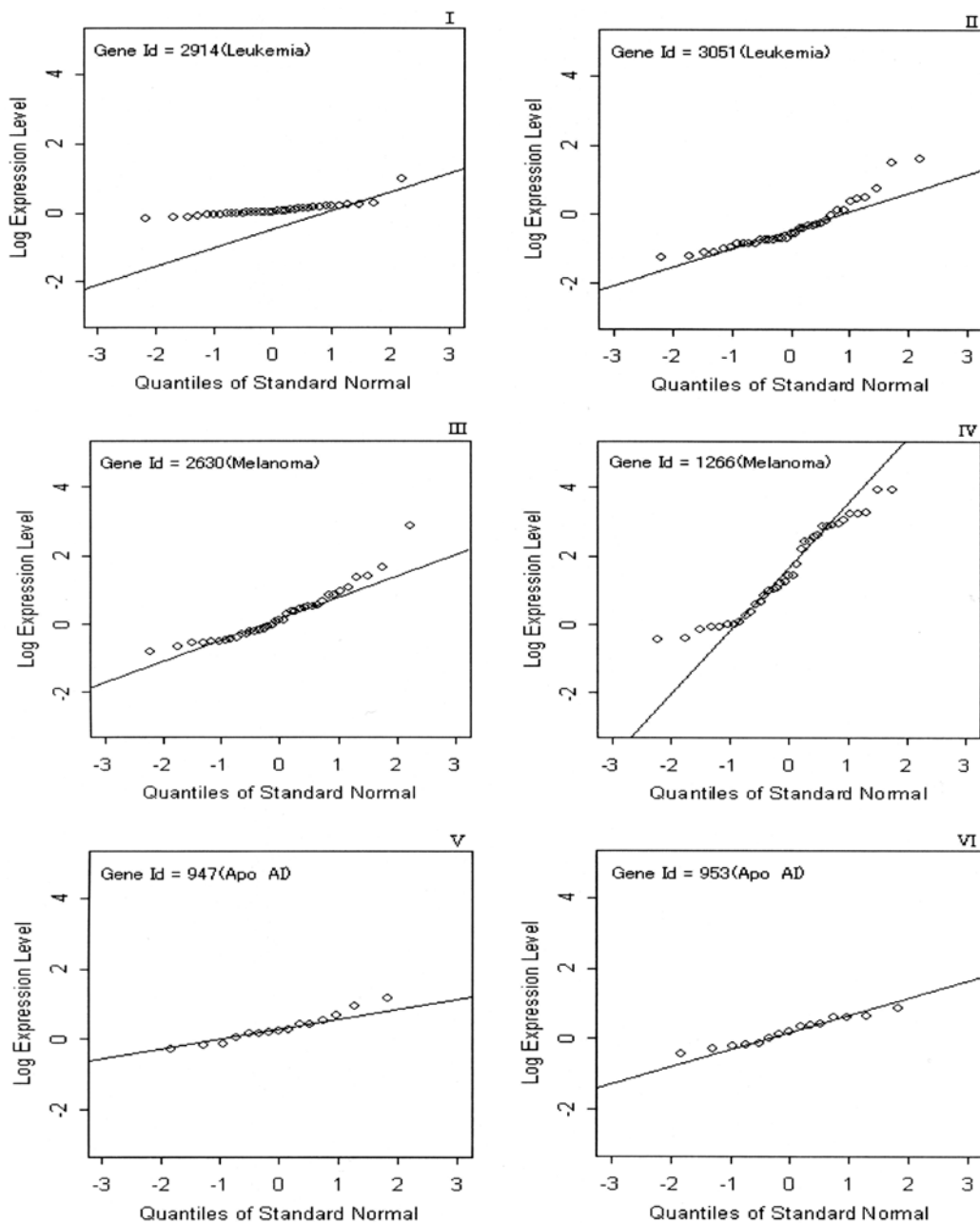


Figure 1 Analysis of leukemia, melanoma and Apo A1 datasets. Plots are the quantile–quantile plots of gene expression levels (in log scale) for checking the normality assumption. Plot I and II are from leukemia data; plot III and IV are from melanoma data, and plot VI and AI are from Apo A1 data.

Table 8 Ranks of top 50 genes selected by the order of absolute values of MMM statistic in leukemia data

Gene	SAM	B-statistic	Samroc	T-statistic	Zhao–Pan	Bayes T-statistic	Fold change	MMM
4535	17	11	25	25	70	507	225	1
2020	83	62	125	108	1562	277	1299	2
804	45	34	59	62	134	55	605	3
5254	143	114	157	174	1826	262	553	4
6281	2	3	1	2	370	4663	218	5
2348	30	17	37	42	200	1374	202	6
1241	462	399	436	491	749	2679	896	7
5087	14	9	21	22	892	1551	1201	8
4328	72	56	92	103	232	605	412	9
1630	423	359	457	521	2776	1274	124	10
4167	127	100	141	151	1185	1484	909	11
2301	1193	1061	1124	1261	2632	2412	6962	12
5772	303	261	338	380	3311	21	414	13
1745	176	148	277	222	1795	1756	480	14
2354	16	10	28	26	613	470	329	15
5191	201	171	235	257	436	272	301	16
3056	164	137	191	213	2159	2994	250	17
5593	392	314	390	452	964	119	860	18
3137	161	132	187	201	1369	916	1170	19
6471	340	344	296	323	2860	531	4237	20
2642	1222	1120	1233	1386	3436	1250	179	21
6283	118	88	128	132	1993	1777	1455	22
2233	6	21	3	4	632	2526	4036	23
6855	235	195	261	282	547	4035	347	24
1306	317	451	250	269	2612	239	6877	25
5625	41	27	49	51	460	173	804	26
3320	11	8	22	21	67	121	839	27
2441	156	133	163	182	1983	4606	789	28
1928	175	149	208	226	1331	1082	1028	29
4546	95	77	97	109	698	124	845	30
4973	745	699	669	754	688	87	581	31
379	721	632	712	816	5363	3804	1144	32
5552	577	507	602	689	979	13	649	33
1144	171	142	189	216	949	2316	275	34
2597	25	14	34	35	129	1188	1897	35
3773	1024	937	925	1037	4406	511	1382	36
3899	58	79	48	50	75	3152	1292	37
6515	269	245	264	289	3802	3617	517	38
5189	1900	2042	1669	1753	6431	4810	6688	39
4823	185	155	206	223	86	205	855	40
1807	573	595	622	529	2326	640	1398	41
1781	1327	1192	1543	1432	1972	25	6888	42
4177	1138	1075	1031	1147	2540	172	4582	43
1703	374	311	398	457	3016	933	380	44
1381	133	104	150	164	1117	340	1077	45
1078	1375	1267	1400	1536	3569	2305	872	46
6084	18	13	23	23	1004	388	403	47
5501	974	892	981	1121	3383	944	818	48
3507	255	241	247	265	1205	967	61	49
6561	100	80	140	125	1023	5	1500	50
Average	373.64	344.66	377.68	407.18	1679.14	1331.54	1442.08	25.5

Note: Only the first four samples from ALL and AML were used.

whereas B-statistic and SAM give the lower average ranks than samroc in both leukemia and melanoma datasets which do not follow the normal distribution. Also, note that B-statistic performs better in the normal case with a large number of samples. As mentioned in Section 2, Smyth¹¹ just conjectured that the B-statistics is not sensitive to the violation of the distributional assumptions, but it was clearly shown by our simulation study. One reason might be that the testing based on B-statistic does not select significant gene by P -value but select by the order of the B-statistic value, and thus it does not depend strongly on the distributional assumptions in the real data analysis.

For small sample case, simulation study shows that Bayes T-test appears to be the best in finding the significant genes for both normal and lognormal cases. In the real data analysis, however, Bayes T-test gives high average rank and thus performs poorly in leukemia and melanoma study and gives the lowest average rank in the Apo AI mouse. Table 7 shows that only Apo AI mouse data follow normal distribution, whereas the other two data do not follow normal distribution, and thus we can find that Bayes T-test is very sensitive to the violation of distributional assumptions. One reason might be that it is sensitive to the choice of hyperparameter ν_0 (pseudo observations) associated with a background variance σ_0^2 in small sample case.¹⁰ In Bayes T-test, for small sample case, the posterior mean estimate is more sensitive to prior mean μ_0 , but the role of prior mean decreases and the estimator is similar to the maximum likelihood estimator as sample size becomes larger. Thus, the testing results can be quite different according to the choice of hyperparameters in small sample case. For small sample case, SAM performs better in the lognormal case, whereas samroc performs better in the normal case. Smaroc still performs well in the unequal variance case when the data follow the normal distribution, but its performance under the non-normal case is not as good as that under the normal case. However, SAM and B-statistic perform well under the lognormal case.

The non-parametric test methods like SAM, samroc, and Zhao–Pan method are usually expected to provide good testing results in small sample case, but our study shows that the performance of Zhao–Pan method is poor in small sample case. To understand why Zhao–Pan method fails in almost all cases, we have conducted an additional simulation study. Zhao–Pan method divides the control group into two subgroups and these two subgroups are used to construct the null statistic z_g , and thus it is worthwhile to see how much the testing result can be affected by the choice of the subgroups. As shown in Table 9, group 1 (control group) consists of 26 arrays, group 1a and group 1b consists of 16 and 10 arrays, respectively, and group 2 (experimental group) consisting of 10 arrays is used to construct the test statistic Z_g . We have generated five different datasets from the

Table 9 Number of significant genes detected by Zhao–Pan method for five different datasets obtained from the original leukemia data (group 1–26 arrays, group 2–10 arrays)

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Group 1a	1:16	1:6,17:26	11:26	1:10, 21:26	5:20
Group 1b	17:26	7:16	1:10	11:20	1:4, 21:26
Group 2	27:36	27:36	27:36	27:36	27:36
Number of detected gene	23	3	10	4	90

original leukemia dataset and counted the numbers of significant genes in each dataset. For example, in dataset 1, 1:16 means that group 1a consists of array 1 to array 16 in the original dataset. We can easily see from Table 9 that the numbers of detected genes are quite different according to the choice of the dataset. Note that three significant genes are detected in the dataset 2, whereas 90 significant genes are detected in the dataset 5. Hence, we can conclude that Zhao–Pan method is quite sensitive to the choice of arrays, and thus it may not be appropriate to construct both null and test statistics using Zhao–Pan method, in general. Especially, in small sample case, testing result seems to be more seriously affected by the choice of the datasets compared to large sample case.

Through our comparison study, we can see that the selection of the significant genes heavily depends on the choice of the testing methods. We can also see that the performance of the testing methods is affected by sample size, distributional assumption, the variance structure and soon. Therefore, to obtain the reliable testing results for detecting significant genes in microarray data analysis, we first need to explore the characteristic of the data and then apply the most appropriate testing method under the given situation.

Acknowledgement

This work was supported by Korea Research Foundation Grant (KRF-2002-075-C00005). JW Lee was also supported by Korea Science and Engineering Foundation Grant (R14-2003-002-01002-0).

References

- 1 Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *The Chipping Forecast* 1999; 21: 33–7.
- 2 Schena M, Shalon, R, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* 1995; 270: 467–70.
- 3 DeRisi L, Penaland L, Brown, PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* 1996; 14: 457–460.
- 4 Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray image. *Biomedical Optics* 1997; 2: 364–74.
- 5 Efron B, Tibshirani R, Strey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; 96: 1151–60.
- 6 Ideker T, Thorsson V, Siehel AF, Hood LE. Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *Journal of Computational Biology* 2000; 7: 805–17.
- 7 Newton MA, Kendzioriski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 2001; 8: 37–2.
- 8 Tusher VG., Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; 98: 5116–21.
- 9 Pan W. A Comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2002; 12: 546–54.
- 10 Lönnstedt I, Speed TP. Replicated microarray data. *Statistical Sinica* 2002; 12: 31–6.
- 11 Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and*

- Molecular Biology* 2004; 3: Article 3: <http://www.bepress.com/Sagmb/vol3/iss1/art3>.
- 12 Dudoit S, Yang YH, Speed TP, Callow MJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; 12: 111–39.
 - 13 Zhao Y, Pan W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2003; 19: 1046–54.
 - 14 Lee MLT, Kuo FC, Witmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* 2000; 97: 9834–9.
 - 15 Kerr MA, Churchill GA. Statistical design and the analysis of gene expression microarray data. *Genetical Research* 2001; 77: 123–8.
 - 16 Long AD, Mangalam HJ, Chan BYP, Tolleri L, Hatfield WG, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *The Journal of Biological Chemistry* 2001; 276: 19937–44.
 - 17 Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001; 17: 509–19.
 - 18 Pan W, Lin J, Le C. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics* 2003; 3: 117–24.
 - 19 Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* 2001; 11: 1227–36.
 - 20 Broberg P. Ranking genes with respect to differential expression. *Genome Biology* 2002; 3: preprint0007.1-0007.23, from <http://genomebiology.com/2002/3/9/preprint/0007>.
 - 21 Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* 2002; 18: 1454–61.
 - 22 Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis, *In Functional Genomics: Methods and Protocols* 2003; 224: 111–36.
 - 23 Golub TR, Slonim DK, Tamajo P, Huard C, Gaosenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531–7.
 - 24 Bittner M, Meltzer P, Chen Y, Jiang, Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Samps N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; 406: 536–40.
 - 25 Darlene RG, Debashis G, Erin MC. Statistical issues in the clustering of gene expression data. *Statistica Sinica* 2002; 12: 219–40.
 - 26 Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* 2000; 10: 2022–29.
 - 27 Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series, B* 2002; 64: 479–98.
 - 28 Arfin SM, Long AD, Ito T, Tolleri L, Riehle MM, Paegle ES, Hatfield GW. Global gene expression profiling in *Escherichia coli* K12: the effect of intergration host factor. *The Journal of Biological Chemistry* 2000; 275: 29672–84.

Copyright of *Statistical Methods in Medical Research* is the property of Arnold Publishers and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.