Full Report

Interdisciplinary Bio Central

Open Access, Open Review Journal

Improved Statistical Testing of Two-class Microarrays with a Robust Statistical Approach

Hee-Seok Oh[1], Dongik Jang[1], Seungyoon Oh[2] and Heebal Kim[2,3,*]

[1]Department of Statistics, Seoul National University, Seoul 151-742, Korea
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea
[3]Department of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea

Subject areas:
Bioinformatics/Computational biology/Molecular modeling, Biomathematics/Mathematical Biology and Medicine

Author contribution: Hee-Seok Oh and Heebal Kim performed research design and writing the paper. Dongik Jang and Sunjin Moon performed programming and data management.

***Correspondence** and requests for material should be addressed to H.B.K. (heebal@snu.ac.kr)

Reviewer: Daehee Hwang, POSTECH, Republic of Korea. Dongsup Kim, KAIST, Republic of Korea.

Editor: Keun Woo Lee, Gyeongsang National University, Republic of Korea.

Received Mar 17, 2010;
Revised Apr 30, 2010;
Accepted May 31, 2010;
Published Jun 01, 2010

Citation: Oh, H-S, et al. Improved Statistical Testing of Two-class Microarrays with a Robust Statistical Approach. IBC 2010, 2:4, 1-6. doi:10.4051/ibc.2010.2.2.0004

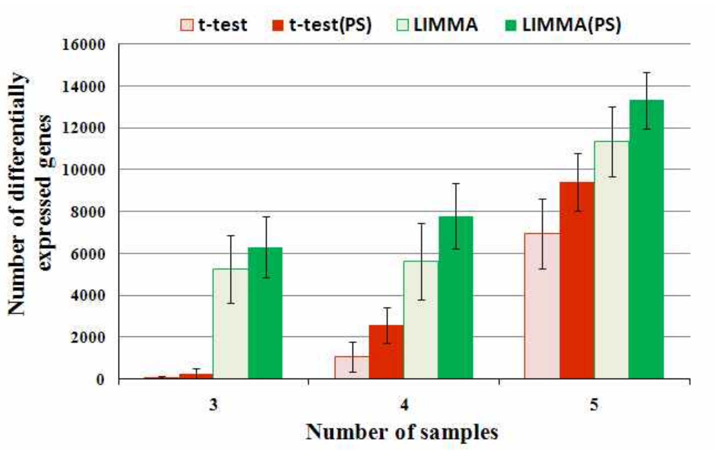
Supporting online materials:
http://www.ibc7.org/article/data_file.php?sid=197&mode=supplemental

Funding: This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-314-C00046) and by the BioGreen 21 program of the Korea Rural Development Administration.

Competing interests: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

The most common type of microarray experiment has a simple design using microarray data obtained from two different groups or conditions. A typical method to identify differentially expressed genes (DEGs) between two conditions is the conventional Student's *t*-test. The *t*-test is based on the simple estimation of the population variance for a gene using the sample variance of its expression levels. Although empirical Bayes approach improves on the *t*-statistic by not giving a high rank to genes only because they have a small sample variance, the basic assumption for this is same as the ordinary *t*-test which is the equality of variances across experimental groups. The *t*-test and empirical Bayes approach suffer from low statistical power because of the assumption of normal and unimodal distributions for the microarray data analysis. We propose a method to address these problems that is robust to outliers or skewed data, while maintaining the advantages of the classical *t*-test or modified *t*-statistics. The resulting data transformation to fit the normality assumption increases the statistical power for identifying DEGs using these statistics.



Keywords: Microarray, *t*-test, empirical Bayes, Pseudo data

Abstract

The performance of the conventional *t*-test for microarray data is considered poor because the level of replication within groups is often low and the expression intensities may not be normally distributed. However, the *t*-test is still one of the most commonly used methods for microarray analysis because of its simplicity, its speed, and its ability to account for the underlying variability in the data. The major drawback in applying the conventional *t*-test and its variations, such as the significant analysis of microarrays (SAM) or an empirical Bayes approach, is that gene expression values of microarray data exhibit a strong departure from the normal distribution. We propose a method to overcome this disadvantage that is robust to outliers or skewed data while maintaining the advantages of each test. Simulation studies showed that the proposed method is more powerful than existing methods, such as the conventional *t*-test, SAM, and the empirical Bayes approach in cases where outliers are present and the data come from skewed distributions. Application on real microarray data showed close agreement with the simulation study. Supplementary information and R software code for generating pseudo data are available at http://snugenome.snu.ac.kr/bootcl.

Introduction

Many scientific papers related to expression profiling using microarray technology have been published since the first miniaturized microarrays for gene expression profiling were reported (Schena et al., 1995). The most common type of microarray experiment has a simple design using microarray data obtained from two different groups or conditions, *e.g.*, cancer and normal tissues (Yoon et al., 2006). A major challenge in this type of experimental setup is the identification of genes the expression of which is significantly different between two conditions (Aittokallio et al., 2003); these are referred to as differentially expressed genes (DEGs). A typical method to identify DEGs between two conditions is the conventional Student's *t*-test (Gosset, 1908). The performance of the *t*-test is considered to be poor as the level of replication within groups for microarray experiments is often low, and the expression intensities may not be normally distributed (Wang and Ethier, 2004). Student's *t*-test, however, is still one of the most commonly used methods for microarray analysis due in part to its simplicity, speed, and ability to account for the underlying variability in the data that makes it superior to simple-minded fold change methods (Papana and Ishwaran, 2006). The *t*-test is based on the simple estimation of the population variance for a gene using the sample variance of its expression levels. As there are problems using the *t*-test with microarray data from a small sample size, alternative statistics have been proposed that share the strength of all genes to obtain a more stable estimate of gene-specific variance, and the resulting statistics are referred to as modified *t*-statistics (Irizarry, 2005). Significance analysis of microarrays (SAM) is a common example of modified *t*-statistics (Tusher et al., 2001). SAM statistics have an added constant ("fudge" factor) to prevent the situation where a small sample variance causes large *t*-statistics. Thus, they are also referred to as penalized or regularized *t*-statistics. Another example of such modified *t*-statistics is based on an empirical Bayes approach available in the Limma R statistical package, which reduces estimated sample variances towards a pooled estimate, producing more stable result when the number of samples is small (Smyth, 2004). The empirical Bayes approach improves on the *t*-statistic by not giving a high rank to genes only because they have a small sample variance (Irizarry, 2005). The basic assumption for this is the equality of variances across experimental groups. The equal variance model across groups has been shown to agree closely with microarray data (Ishwaran and Rao, 2003, Ishwaran and Rao, 2005). The *t*-test suffers from low statistical power because of the assumption of normal and unimodal distributions for the microarray data analysis. Variations of the conventional *t*-test, SAM, and the empirical Bayes approach available in the Limma R statistical package, are also not free from the problem that gene expression values of microarray data exhibit significant departures from the normal distribution (Yan et al., 2005). We propose a method to address these problems that is robust to outliers or skewed data, while maintaining the advantages of the classical *t*-test or modified *t*-statistics. The resulting data transformation to fit the normality assumption increases the statistical power for identifying DEGs using these statistics.

Materials and Methods

Problem statement

Let x_{ij} be the log-expression values for a gene transcript *i* from microarray chip *j*, where *i* = 1, 2, …, *m* and *j* = 1, 2, …, *n*. Define a standardized expression level of the *i*th gene for a specific *j*th sample as

$$u_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where $\mu_j = \sum_{i=1}^{m} \frac{x_{ij}}{m}$ and $\sigma_j = \sqrt{\sum_{i=1}^{m} \frac{(x_{ij} - \mu_j)^2}{m-1}}$. The conventional *t*-test has been used to search for differences in gene expression as follows. We now consider a two-sample problem. Suppose that we collect the standardized expression levels of the *i*th gene from a normal cell and an abnormal cell. The *t*-test to identify the differential effect for a specific gene *i* is

$$t_i^{(1)} = \frac{\bar{u}_{i,+} - \bar{u}_{i,-}}{\sqrt{\frac{\hat{\sigma}_{i,+}^2}{n_1} + \frac{\hat{\sigma}_{i,-}^2}{n_2}}} \qquad (2.1)$$

where $\bar{u}_{i,+}, \bar{u}_{i,-}$ are the sample means of u_i for two groups (*e.g.*, cancerous and normal tissue), n_1 and n_2 denote the sample size for the group, and $\hat{\sigma}_{i,+}^2$, $\hat{\sigma}_{i,-}^2$ are the sample variance for the two groups. In the case of equal variance, $\sigma_{i,+}^2 = \sigma_{i,-}^2$, the test statistic of the *t*-test becomes

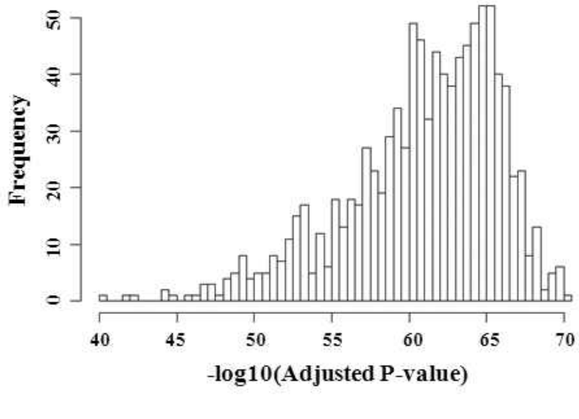


Figure 1. Frequency histogram of the *P*-value using the Shapiro-Wilk test for normality. The *y*-axis represents the frequency of probes for a given *P*-value bin. The expression value distributions of each probe were drawn from 5000 samples randomly selected from the Affymetrix GeneChip Human Genome U133 Plus 2.0 array in GEO (http://www.hcbi.nlm.nih.gov/geo/), accession number GPL570. The normality tests were performed using 1000 randomly selected probes.

$$t_i^{(2)} = \frac{\bar{u}_{i,+} - \bar{u}_{i,-}}{\hat{\sigma}_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (2.2)$$

where $\hat{\sigma}_i^2$ denotes the pooled estimate of variance,

$$\hat{\sigma}_i^2 = \frac{(n_1 - 1)\hat{\sigma}_{i,+}^2 + (n_2 - 1)\hat{\sigma}_{i,-}^2}{n_1 + n_2 - 2}.$$

The above procedure is required to ensure that the data u_{ij} are normally distributed. The normality of real gene expression values was assessed using the Shapiro-Wilk test on 5000 microarray samples selected randomly from a commonly used microarray platform. None of the distributions of expression values for 1000 randomly selected probes showed normality across the 5000 microarray samples (Figure 1). Real data tend to have outliers or to be skewed.

For this reason, the conventional *t*-test could not be used directly for identifying the differences of gene expression. We propose a method that is robust to outliers or skewed data, while maintaining the advantages of the classical or modified *t*-tests. The key element of the proposed method is to generate pseudo data that can be simply transformed to being very closely normally distributed. We then apply the conventional *t*-test, SAM test, and an empirical Bayes approach available in the Limma R statistical package to the pseudo data to compare the gene expression differences. As shown in the results section, all the tests increased the statistical power in both the simulated and real data analysis.

Gaussian Distribution and Sample Mean

Given a set of observations $\{y_1, y_2, \cdots, y_n\}$, an estimator T_n of location parameter θ can be defined as the minimizer of an expected loss

$$\underset{\theta}{\operatorname{argmin}} \int \rho(y - \theta) dF_n(y) \equiv \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \theta),$$

where F_n is the empirical distribution function of the observations $\{y_1, \cdots, y_n\}$. Therefore, if $\rho(x)$ is differentiable with an absolutely continuous derivative $\psi(x)$, then the estimator of the location parameter that solves the equation is

$$\sum_{i=1}^{n} \psi(y_i - \theta) = 0 \qquad (3.1)$$

Now consider the case of $\rho(x) = x^2$ that corresponds to $-\log f(x)$, where $f(x)$ denotes the density of standard Gaussian distribution. Eq. (3.1) can be written as

$$\sum_{i=1}^{n} 2(y_i - \theta) = 0 \qquad (3.2)$$

that provides a sample mean $\bar{y}_n = \sum_{i=1}^{n} \frac{y_i}{n}$. This indicates that the sample mean is an optimal choice when the difference $\varepsilon = y - \theta$ follows the Gaussian distribution. That is, if ε does not come from a Gaussian distribution, the sample mean may not be reasonable for representing the location parameter.

Pseudo Data

Some values of $\varepsilon_i = y_i - \theta$ will be very large if outliers are present, and the sample mean from Eq. (3.2) will be markedly affected by these outliers. To overcome this problem, it is natural to use a different $\psi(\varepsilon)$, which is the same as 2ε for $|\varepsilon| \le c$, and is negligible when $|\varepsilon| > c$. Then, the result from Eq. (3.1) may be robust to outliers. A simple choice of $\psi(x)$ displayed in Figure 2(a) is

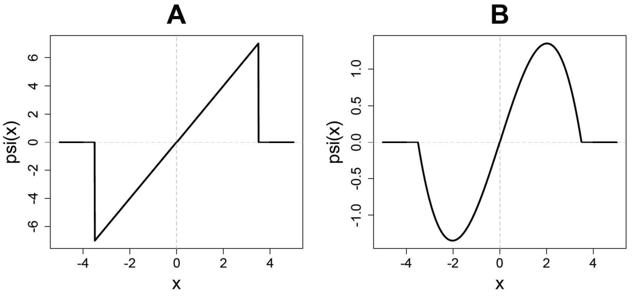


Figure 2. Examples of $\psi(x)$: (a) simple choice of $\psi(x)$, and (b) Tukey's biweight function.

$$\psi_H(x) = \begin{cases} 2x & \text{f} \quad |x| \le c \\ 0 & \text{ot herwise} \end{cases}.$$

Here, c is a cutoff point that is typically chosen to be $c = \mathcal{K}\sigma$. The cutoff point was fixed at $c = \mathcal{K}\, MAD$ in this study, where MAD is the median absolute deviation defined as MAD = 1.4826 × median(|x-median(x)|), and the constant $\mathcal{K}$ was chosen considering the robustness of the estimate as well as the efficiency at the normal distribution. More specifically, large values of $\mathcal{K}$ mean that the efficiency of the estimate increases, but maintaining the robustness becomes more difficult. In this study, we used three values of $\mathcal{K}$ (3.5, 3.8, and 4.0) that are sufficient to cover most of the range of the Gaussian random variable x with a mean of 0. Another example of $\psi(x)$ is Tukey's biweight function (Figure 2(b)),

$$\psi(x) = \begin{cases} x\left[1 - \left(\frac{x}{c}\right)^2\right] & \text{f} \quad |x| \le c \\ 0 & \text{ot herwise} \end{cases}.$$

Let us define pseudo data as

$$\tilde{y}_i = \theta + \frac{\psi(y_i - \theta)}{2} \qquad (4.1)$$

Thus, the pseudo data can be considered as a transformation with bounded errors so there are no outliers. The concept of pseudo data was introduced by Huber (Huber, 1973) and Cox (Cox, 1983) derived the asymptotic linearization of *M*-type smoothing splines based on pseudo data. More recently, Oh *et al.* (Oh et al., 2007) generalized Cox's results to a general class of roughness penalties. To illustrate the idea of pseudo data explicitly, we consider the problem of obtaining the sample mean of pseudo data $\tilde{y}_i$ that requires solving the equation

$$\sum_{i=1}^{n} 2(\tilde{y}_i - \theta) = 0 \qquad (4.2)$$

From the definition of pseudo data, the above equation is equivalent to Eq. (3.1) with original values y_i such that

$$\sum_{i=1}^{n} \psi(\tilde{y}_i - \theta) = 0 \qquad (4.3)$$

We know from the derivation above, that although the original data contain some outliers, the sample mean computed by the pseudo data is a good representation of the true location parameter as the pseudo data tend to be Gaussian without outliers. Hence, we can use the pseudo data with the conventional *t*-test, SAM, and the empirical Bayes approach to compare the gene expression levels of two groups when outliers are present.

Results & Discussion

Empirical Pseudo Data and Practical Algorithm

Here, we used a practical algorithm for identifying the difference of gene expression level in two distinct groups. The proposed algorithm was formed by combining the concept of pseudo data with traditional *t*-tests. In practice, θ is unknown and thus the pseudo data of Eq. (4.1) are not available. Instead, we consider a fixed point analogy to pseudo data. If $\hat{\theta}$ is an estimate of θ, we form the empirical pseudo data

$$z_i = \hat{\theta} + \frac{\psi(y_i - \theta)}{2} \qquad (5.1)$$

As the value of z_i depends on the estimate $\hat{\theta}$, we obtain the empirical pseudo data z_i in the following iterative manner.

Starting with initial estimate $\hat{\theta}^0$, repeat over *l* until convergence occurs:

Step 1: Form z_i^l from Eq. (5.1) and
Step 2: Compute $\hat{\theta}^{l+1}$ based on the values of z_i^l.

Note that we use the sample median or mean for $\hat{\theta}^l$. We generated 200 random variables from the *t*-distribution with 3 degrees of freedom as shown in Figure 3 to show the performance of empirical pseudo data. The empirical pseudo data seemed to follow the normal distribution even though the original observations did not.

The proposed algorithm can be summarized as based on the derivation above: (1) obtain the converged empirical pseudo data from two groups and then compute the corresponding u_{ij}, and 2) run the *t*-test, SAM, and Limma test with u_{ij}. Note that the above procedure can be easily used for identifying the differences in gene expression under several conditions (tissues) using a correction method for multiple tests, such as that described by Benjamini and Hochberg (Benjamini and Hochberg, 1995)

Simulation Study

This section investigates the practical performance of the proposed method. We compared the proposed robust method to these existing methods: (i) the conventional *t*-test of (2), (ii) the proposed robust method with $\psi_H(x)$ and *c*=3.8 and sample variance in the second step, (iii) Limma reported by Smyth (Smyth, 2004), and (iv) SAM reported by Tusher *et al.* (Tusher et al., 2001). Note that the multiple test correction of Benjamini and Hocberg (15) was used for comparison when the proposed method was implemented. In addition, the sample median was used for the initial estimate $\hat{\theta}^0$, and the sample means were used for $\hat{\theta}^l$. Now, we generate artificial data that represent real data with 10,000 genes and 10 chips. We create data based on a gene platform to integrate realistic gene expression data. First, we define the null set as

$$m_{\downarrow ij} = \left(\max_{\downarrow i} - \min_{\downarrow j}\right) X_{\downarrow ij} + \mu_{\downarrow ij} + \varepsilon_{\downarrow ij} \qquad (6.1)$$

where X_{ij} are generated from distribution $\frac{Q_{ij}}{\sqrt{\sum_{i=1}^{10} Q_{ij}}}$, $Q_{ij} \sim N(0,1)$ and

$$\varepsilon_{ij} \sim N\left(0, \sigma_j^2\right), t_{\frac{2\sigma_j^2}{\sigma_j^2 - 1} I_{\sigma_j^2 > 1} + 5\, I_{\sigma_j^2 \le 1}} \text{ and } \chi_{\frac{\sigma_j^2}{2}}.$$

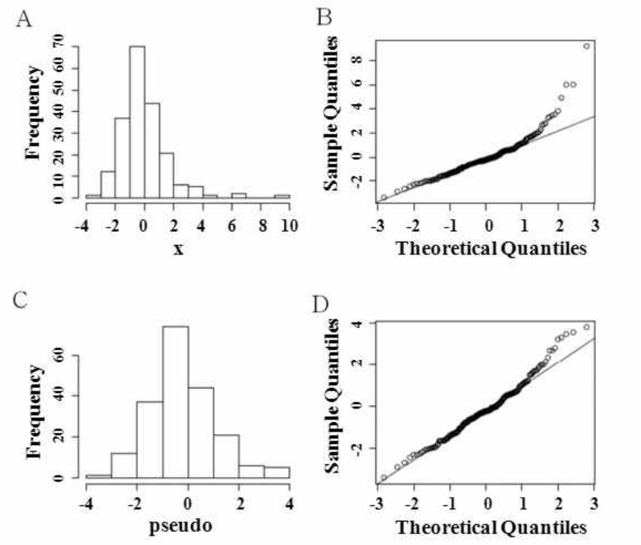


Figure 3. Performance of the empirical pseudo data, from left to right, top to bottom: histogram of the original observation and its Q-Q plot, and histogram of the empirical pseudo data and its Q-Q plot.

Here, $\min_{\downarrow j}, \max_{\downarrow j}, \mu_{\downarrow j}$ and σ_j^2 denote the minimum, maximum, mean, and variance of the gene, respectively, which are selected randomly in the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array GPL570 platform in GEO. That is, m_{ij} represent real genes with three different types of independent and identical errors from the standard Gaussian distribution, the *t*-distribution and χ^2 distribution, respectively. Note that the setting of generation X_{ij} was used by Cui *et al.* (Cui et al., 2005). The last two types represent the cases in which outliers are present and the data come from a skewed distribution. We calculate the empirical power that is defined in the following steps for assessing the performance of each method:

(1) Randomly select a gene in the gene platform and compute the minimum, maximum, mean, and variance of the selected gene,
(2) Generate data from the set of X_{ij} according to distribution $\frac{Q_{ij}}{\sqrt{\sum_{i=1}^{10} Q_{ij}}}$,
(3) Generate data from the null set of m_{ij} ,
(4) Obtain data to represent alternatives shifting by a value of a specified distance,
(5) Apply each method to the data generated in steps (1) and (2),
(6) Compute the adjusted *P*-value, and
(7) Calculate the empirical power where

$$em\ piical \quad pow\ er = \frac{\#\ of\ p_{value} < 0.05}{n}$$

after repeating steps (1)–(4) *n* times. Figure 4 shows the empirical power of each method according to different distance specified in step (2).

We obtain the following empirical observations from the simulation results: (i) The empirical power of the proposed method is comparable or superior to those of the classical *t*-test, Limma, and SAM for the Gaussian error case. (ii) The proposed method outperforms the classical *t*-test, Limma, and SAM for the non-Gaussian case. (iii) For all cases, the proposed robust approach provides the best performance when the distance is not large.

Overall, the simulation results suggest that the proposed method is preferable for identifying differences in gene expression data that come from various distributions. Note that the results of the proposed method with different values of $\mathcal{K}$ and initialization are similar, and are thus omitted. Figure 4 shows the simulation results for empirical power as a function of distance with 10 replications for each distribution type with sample sizes of 3, 6, and 9. In all cases, the empirical powers using the pseudo sampling method were superior to the power of each statistic without pseudo sampling. This was especially so for the non-Gaussian case and a small sample size. We also examined the type-I error rate for each case with a set of simulated data without a given distance. The type-I error rates were similar for the tests with and without pseudo sampling, and were sufficiently close to zero to be negligible.

Performance on Microarray Data

Real data sets used in this study consisting of raw data of ten ovarian endometriosis and ten matched control endometrium (Hever et al., 2007) were downloaded from the Gene Expression Omnibus (GEO) database (http://www.hcbi.nlm.nih.gov/geo/) (accession no. GSE7305). From the ovarian endometriosis and control endometrium samples, *N* (=3,4,5) samples were randomly selected 10 times, and compared using the ordinary *t*-test, Limma, *t*-test (ps-data), and Limma (ps-data) with a significance threshold of *P*<0.05. The probes that were identified as significantly up- or down-regulated were counted for each statistical model. The real data analysis showed results similar to the simulation study, *i.e.*, each statistical test showed superior statistical power when it was combined with the pseudo sampling method (Figure 5). In ascending order, the empirical power of the tests seemed to be: ordinary *t*-test, *t*-test with pseudo sampling, Limma test, and Limma test with pseudo sampling. The empirical power result of the real data analysis agrees with the simulation result. Thus, in general, the Limma package combined with our pseudo sampling algorithm provided the best empirical power. The real microarray data used here was such that many DEGs could be identified with small sample sizes. However, many researchers have experienced frustration in not finding any DEGs in their microarray data. Replication is a straightforward method for improving the quality of inferences made from experimental studies, and is clearly necessary in microarray experiments. A trade-off between cost and statistical power frequently arises in gene expression microarray experiments because such experiments are costly and involve RNA samples that are often difficult to obtain (Pavlidis et al., 2003). It is recommended that a minimum of 5 biological cases per group be analyzed (Allison et al., 2006) for designs in which two groups of cases are evaluated for differential expression. This sample size is considered a minimum, not an optimum, and applies only to differential expression testing for two groups of cases, not classification (Allison et al., 2006, Pavlidis et al., 2003, Tsai et al., 2003). However, there have been a number of reports based on less than 5 biological samples per group. Therefore, increasing the statistical power for microarray data analysis is crucial not only for two-group studies, but also for comparison of several groups and time-series data analysis.

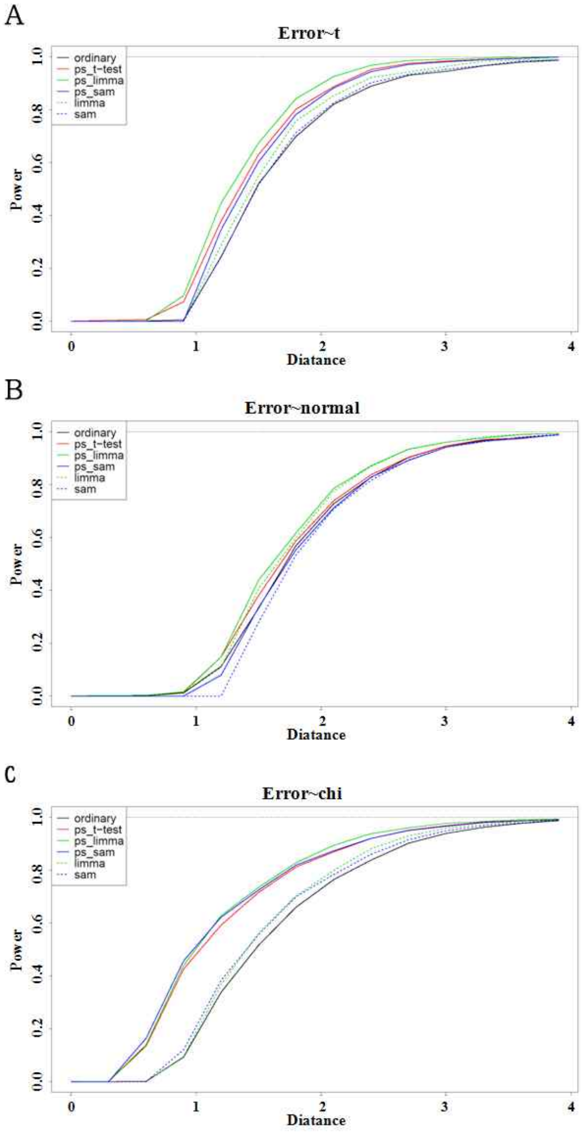


Figure 4. Empirical power as a function of distance for each method with (a) Gaussian distribution, (b) *t*-distribution, and (c) λ^2-distribution. “Ordinary” and “ps” indicate ordinary *t*-test and pseudo data approach, respectively.

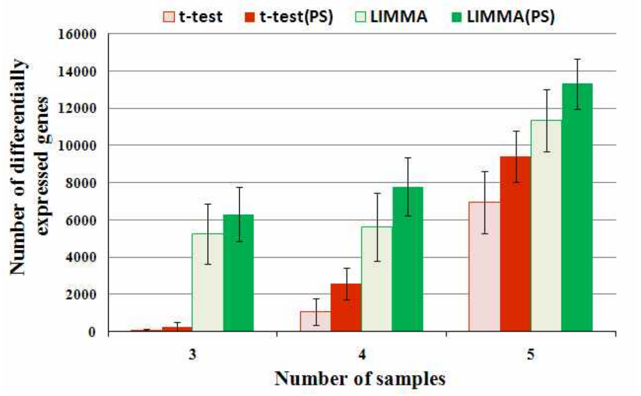


Figure 5. Comparison of statistical powers for real microarray data.

Conclusion and Prospects

We propose a method to address these problems that is robust to outliers or skewed data, while maintaining the advantages of the classical *t*-test or modified *t*-statistics. The resulting data transformation to fit the normality assumption increases the statistical power for identifying DEGs using these statistics. The proposed algorithm has several advantages. First, it can be easily implemented. Second, it can be extended to the analysis of variance for comparing gene expression levels of several groups simultaneously. Third, the proposed method can be applied to a quantile comparison of gene expression levels that provides a more complete view of the statistical data landscape.

Acknowledgments

This work was supported by the Korea Research Foundation

Grant funded by the Korean Government (KRF-2008-314-C00046) and by the BioGreen 21 program of the Korea Rural Development Administration.

References

Aittokallio, T., Kurki, M., Nevalainen, O., Nikula, T., West, A. and Lahesmaa, R. (2003). Computational strategies for analyzing data in gene expression microarray experiments. *J Bioinform Comput Biol* **1**, 541-586.

Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55-65.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289-300.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *Ann. Statist* **11**, 530-551.

Cui, X., Hwang, J. T., Qiu, J., Blades, N. J. and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59-75.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika* **6**, 1-25.

Hever, A., Roth, R. B., Hevezi, P., Marin, M. E., Acosta, J. A., Acosta, H., Rojas, J., Herrera, R., Grigoriadis, D., White, E., Conlon, P. J., Maki, R. A. and Zlotnik, A. (2007). Human endometriosis is associated with plasma cells and overexpression of B lymphocyte stimulator. *Proceedings of the National Academy of Sciences* **104**, 12451-12456.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1**, 799-821.

Irizarry, R. A. (2005). From CEL files to annotated lists of interesting genes. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor?Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, eds*, 434-435.

Ishwaran, H. and Rao, J. S. (2003). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal of the American Statistical Association* **98**, 438-456.

Ishwaran, H. and Rao, J. S. (2005). Spike and Slab Gene Selection for Multigroup Microarray Data. *Journal of the American Statistical Association* **100**, 764-781.

Oh, H. S., Nychka, D. W. and Lee, T. (2007). The Role of Pseudo Data for Robust Smoothing with Application to Wavelet Regression. *Biometrika* **94**, 893.

Papana, A. and Ishwaran, H. (2006). CART variance stabilization and regularization for high-throughput genomic data. *Bioinformatics* **22**, 2254-2261.

Pavlidis, P., Li, Q. and Noble, W. S. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics* **19**, 1620-1627.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 1027.

Tsai, C. A., Hsueh, H. M. and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* **59**, 1071-1081.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121.

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

Yan, X., Deng, M., Fung, W. K. and Qian, M. (2005). Detecting differentially expressed genes by relative entropy. *J Theor Biol* **234**, 395-402.

Yoon, S., Yang, Y., Choi, J. and Seong, J. (2006). Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics* **22**, 2898-2904.