Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Andrew et al., J Biom Biostat 2015, 6:5
DOI: 10.4172/2155-6180.1000265

Research Article Open Access

Methods for Identifying Differentially Expressed Genes: An Empirical Comparison

Andrew H, Florence G and Golam Kibria BM*

Department of Mathematics and Statistics, Florida International University, Miami, Florida, USA

Abstract

Microarray technology, which observes thousands of gene expressions at once, is one of the popular topics in recent decades. When it comes to the analysis of microarray data to identify differentially expressed (DE) genes, many methods have been proposed and modified for improvement. However, the most popular methods such as Significance Analysis of Microarrays (SAM), samroc, fold change, and rank product are far from perfect. In order to determine which method is most powerful, it comes down to the characteristics of the sample and distribution of the gene expressions. The most practiced method is usually SAM or samroc but when the data tends to be skewed, the power of these methods decreases. With the concept that the median becomes a better measure of central tendency than the mean when the data is skewed, the test statistics of the SAM and fold change methods are modified in this paper. This study shows that the median modified fold change method improves the power for many cases when identifying DE genes if the data follows a lognormal distribution.

Keywords: Microarray technology; Lognormal distribution; Expressed genes

Introduction

Microarray technology has allowed researchers to observe thousands of gene expressions all at once. Gene expression in cells is of relevance because it allows a way to pinpoint disease markers that are related to medical treatments [1]. A job that many researchers may want to perform would be to identify which genes in a cell are differentially expressed. For example, a researcher may need to conduct an experiment to discover differentially expressed genes between two experimental conditions. For explanation purposes this could be between healthy patients and patients who have a condition of interest such as cancer. Microarray analysis will allow the researcher to find which genes are expressed differently between these two groups of patients. The researchers will then be able to develop a treatment that targets these specific genes and create a more effective type of therapy. Further information on microarray technology can be found in Majtan et al. [2].

Over the years many methods have been studied to perform the analysis of microarray data. These methods can be categorized into two types, parametric methods and nonparametric methods. Examples of parametric methods are the t-test, Bayes t-test [3], an analysis of variance approach, and the B-statistic method. Nonparametric methods, on the other hand, have become very attractive in this field of research because of the previous costs of microarray experiments and the availability of replicated data has made it difficult to obtain large samples. Nonparametric methods include Significance Analysis of Microarrays (SAM) proposed by Tusher et al. [4], samroc, which uses a very similar test statistic to SAM's in addition to the use of a receiver operating characteristic (ROC) curve [5], the mixture model method (MMM) [6], nonparametric empirical Bayes method [4] and the Zhao-Pan method.

A variety of comparisons between methods have been performed in the past to find which method is most reliable in discovering true differentially expressed genes. The main purpose in these comparisons is to find the method that correctly identifies the highest proportion of the true differentially expressed (DE) genes as DE while maintaining a small proportion of equivalently expressed (EE) genes being falsely identified as DE.

One of the most widely used methods for microarray analysis is the previously mentioned SAM. However, SAM is not a completely robust method and some shortcomings arise. Many researchers have attempted to modify the method in order to make it more reliable. When the number of significant genes is fairly large in a data set, the estimated number of significant genes by SAM is affected and the test is less powerful. As a solution, Pan et al. [6] suggested the use of MMM to estimate the distribution of the null and test statistic. The MMM allows for identifications of a rejection region for any type 1 error rate. In another attempt to fix this bias, Van de Wiel proposes a method using rank scores within SAM. Just by replacing the data with rank scores, the tendency of SAM to produce a biased estimate of DE genes is eliminated. The results are only valid though when the number of samples, N, is not "too small". On the basis of the test statistic used in SAM, Broberg's [5] created the samroc method. Broberg found that when the number of DE genes is large, then the samroc method is likely to work better than SAM. However, in most of the tests performed, the two methods worked just as well as each other when samroc did not outperform SAM.

Breitling et al. [7] adopted another approach to identify differentially expressed genes called rank product in an attempt to exceed SAM. The results showed that, while being a simpler method than SAM, rank product outperformed SAM in identifying DE genes, even with very small data sets. It is also seen that the rank product method performed very similarly to fold change. Fold change (FC) is a popular method often used because of its simplicity and easy understanding.

***Corresponding author:** Golam Kibria BM, Professor, Department of Mathematics and Statistics, Associate Editor-Communications in Statistics A and B, Coordinating Editor-JPSS, Florida International University, Miami, Florida, USA, Tel: (305) 348 1419; E-mail: kibriag@fiu.edu

Received October 16, 2015; **Accepted** November 16, 2015; **Published** November 25, 2015

Citation: Andrew H, Florence G, Golam Kibria BM (2015) Methods for Identifying Differentially Expressed Genes: An Empirical Comparison. J Biom Biostat 6: 265. doi:10.4172/2155-6180.1000265

Comparisons across methods are interesting because each method usually results in outcomes without much agreement. In Jeffery et al. [8] it is found that only 8 to 21% of the genes are commonly identified between the ten different methods being compared including SAM, samroc, fold change, and rank product. The study shows that many factors such as number of genes and number of samples influences which method will obtain the best result. It is concluded that rank product works well under settings with low number of samples and the ROC curve performed well under data sets with large sample sizes. The conclusion by Kim et al. [9] is similar to that of Jeffery et al. [8], noting that the sample size, distribution, and equal variance assumptions of each test greatly impact which test performs better. Our study shows that SAM outperformed samroc when the data follows a lognormal distribution.

Despite the advancement of next generation sequencing (NGS) as an alternative to microarrays, research in analysis of microarrays is still very relevant. Researchers in labs are more comfortable and confident with using microarrays as the technology has been around for a long time and it is less complicated than NGS [10]. Figuring out the most efficient method to identify differentially expressed genes under particular data settings can help master the data analysis step in microarray research.

The focus of the present study is a comparison of the top performing and popular methods SAM, samroc, rank product, and fold change along with modified versions of the SAM method and the fold change rule. As it is evident in Kim et al. [9] and Jeffery et al. [8], sample size and distributional assumption of the data largely impacts the decision of which is the superior method to choose when identifying differentially expressed genes. The aim of this paper was found after evaluating previous research and understanding the biggest drawbacks in this area. Several settings of lognormal cases with various sample sizes will be tested under each of the methods. For the first time, a modification that uses median in place of the mean in the test statistics of SAM and the fold change rule will be made in this paper. The modifications follow from the concept that the median is a better measure of central tendency than the mean when describing skewed data. The expectation is that using the median will better represent the average gene expressions when the microarray data follows a skewed distribution. The modification to fold change will be shown to improve results in identifying differentially expressed genes under skewed data settings. A table of cutoff values for fold change and its modified version is also included in the present study.

The organization of this paper is as follows. In section 2, the statistical techniques are given. A simulation study under the different settings of sample size and skewness is performed on each of the methods in section 3, section 4 will include the application and analysis of a real data set. Finally, conclusions will be made along with a statement of some concerns and future possible research in section 5.

Statistical Methods

This section is a review of several favored statistical methods for identifying differentially expressed genes in microarray datasets. The performance of the methods on data that follow a lognormal distribution are of interest. Let the i^th^ gene expression level of the j^th^ sample under condition 1 be represented by X_{ij} and the i^th^ gene expression level of the k^th^ sample under condition 2 be represented by Y_{ik}, where j = 1,…,J, k = 1,…,K, which represents replicates under condition 1 and 2 respectively. The gene number is represented by i, where i = 1,…,n. For this study n = 5000 genes. The number of genes, n, was chosen to be 5000 based on the work of Schwender et al. research.

SAM

The test statistic in SAM is very similar to the test statistic from the simple t-test. The difference lies on the introduction of a small constant, s_0, in the denominator. The test statistic for SAM is as follows:

$$d(i) = \frac{\overline{X}_i - \overline{Y}_i}{s(i) + s_0} \tag{2.1}$$

where X_i is the expression of the i^th^ gene under experimental condition 1 and Y_i is the expression of the i^th^ gene under experimental condition 2 (i = 1,…,n). Further, $\overline{X}$ and $\overline{Y}$ are the mean expression levels under conditions 1 and 2 respectively for gene i.

The "gene-specific scatter" or standard deviation s(i) is defined:

$$s(i) = \sqrt{\frac{\frac{1}{J} + 1/K}{J+K-2} \cdot \left\{ \sum_{j=1}^{J} (X_{ij} - \overline{X}_i)^2 + \sum_{k=1}^{K} (Y_{ik} - \overline{Y}_i)^2 \right\}} \tag{2.2}$$

where J is the number of replicates in experimental condition 1 and K is the number of replicates in experimental condition 2.

The constant, s_0, is added in order to correct the issue that the traditional t-test faces. The problem with the t-test occurs when genes have low expression levels and yield a small sample variance. The combination of those two factors lead to producing a large test statistic making it very likely that the gene will be identified as DE. The value of s_0 represents a percentile of the standard deviation values of all the genes. The method to compute this value can be found on Page 30 of the SAM user guide [1].

In order to find which genes are DE, SAM calls an algorithm to create the null scores by pooling the data together across the two treatments per gene B times, where B is the total number of permutations. For each permutation, SAM finds the null statistic by using the same formula as the original test statistic, resulting in a total of B null statistics for each gene. The mean of the null statistic is then found for each gene and plotted against the ordered test statistic. The absolute differences between the two values are then found and compared against a cutoff value to determine whether or not there is a significant difference [4]. The cutoff value can be obtained by following the method explained on Page 29 of the SAM user guide [1].

Samroc

Broberg's [5] approach to identifying lists of significant genes while minimizing the rate of false positives and false negatives consists of ranking genes in order of likelihood of being differentially expressed. The test statistic is similar to that of SAM, however the constant s_0, is chosen in a different manner [9].

Fold change

According to McCarthy and Smyth [11], the earliest publications in analyzing microarray data to identify differentially expressed genes used the fold change rule. The fold change rule is defined as follows [9]:

$$FC_i = \frac{\max\left(\overline{X}_i, \overline{Y}_i\right)}{\min\left(\overline{X}_i, \overline{Y}_i\right)} \tag{2.3}$$

where $\overline{X}$ and $\overline{Y}$ are the mean expression levels under conditions 1 and 2 respectively for gene i. The typical accepted cutoff value for the fold change rule is $FC_i > 2$ [11]. McCarthy and Smyth also mention that a disadvantage of the fold change rule is that it does not take variability into consideration. Since it does not account for variability, it makes it difficult to make sense of a set cutoff value. The shortfalls of the fold change rule led to the development of more sophisticated tests such as SAM, however they also have their flaws and do not have the intuitive appeal which the fold change rule has [7].

Rank product

The rank product method was created with overcoming the problems of fold change in mind, while being statistically rigorous and simple at the same time [7]. After the rank product method gained popularity as a method to detect differentially expressed genes in microarray data, Koziol [12] extended the process to a two sample setting. Koziol defines the test statistic as follows:

$$RP_i = \prod_{j=1}^{J} R_{ij}^{1/J} \div \prod_{k=1}^{K} R_{ik}^{1/K} \tag{2.4}$$

where J is the number of replicates in experimental condition 1, K is the number of replicates in experimental condition 2, and the rank is taken among the expressions in a single sample, across the n genes, for each sample. R_{ij} represents those ranks assigned to the ith gene under condition 1 and R_{ik} will be those ranks assigned to the ith gene under condition 2. Further, the monotone log transformation is taken on the test statistic to obtain a better approximation of the null distribution and the resulting statistic is:

$$\log\left(RP_i\right) = 1/J\sum_{j=1}^{J}\log\left(R_{ij}\right) - 1/K\sum_{k=1}^{K}\log\left(R_{ik}\right) \tag{2.5}$$

According to Koziol [12], "the exact distribution of log(RP_i) can be tedious" so a normal approximation of the distribution should be adequate, especially for large samples. If there is skewness in the data, then this approximation may not be adequate.

Median fold change

It has been shown that microarray data is consistent and well approximated by the lognormal distribution [13]. The lognormal distribution is known to be a skewed distribution and the best measure of central tendency for this type of distribution is the median [13,14].

With the prevailing use among biologists as seen in [13] because of its attractive nature and simplicity, we are proposing the following modification to the fold change rule:

$$FC_{Mi} = \frac{\max(\tilde{x}_i, \tilde{y}_i)}{\min(\tilde{x}_i, \tilde{y}_i)}. \tag{2.6}$$

Instead of using the average expression levels of the ith gene under condition 1 and 2, $\bar{X}_i$ and $\bar{Y}_i$, when calculating the fold change, the median expression levels for the ith gene, $\tilde{X}_i$ and $\tilde{Y}_i$ under each condition is used.

$$\tilde{X}_i = \text{median}(X_{i1}, X_{i2}, \ldots\ldots. X_{iJ}) \tag{2.7}$$

$$\tilde{Y}_i = \text{median}(Y_{i1}, Y_{i2}, \ldots\ldots Y_{iK}) \tag{2.8}$$

Simulation Study

Since a theoretical comparison among the test statistics is not possible, a simulation study has been conducted to compare the performance of the test statistics in this chapter. In this section, the performance of SAM, samroc, fold change, rank product and the proposed modifications of fold change using median are compared by applying the methods to simulated gene expression data sets. The methods are compared under the case where the data is simulated to follow a lognormal distribution. The simulations of several combinations of sample sizes have been done while also using three different levels of skewness, slight, moderate, and high.

Simulation techniques

The simulation is performed by generating 5000 genes where 500 of them are knowingly differentially expressed. A matrix, W, is generated of size (5000 x (J + K)), J is the number of samples from condition 1 and K represents the number of samples from condition 2. As stated earlier, each data point in the matrix represents a gene expression, X_{ij} and Y_{ik}. The ith gene expression level under condition 1 is represented by X_{ij} and the ith gene expression level under condition 2 is represented by Y_{ik}.

The comparison between SAM, samroc, fold change, rank product, and the proposed modifications using median are performed under cases of randomly generated data from the lognormal distribution. Different levels of skewness are considered: slightly, moderately, and highly skewed. The levels of skewness will be implemented by setting σ = 1,1.2,1.5 respectively. The data follows the model:

$$X_{ij}\begin{cases}\eta_{ij} & \text{if } \ 1 \le \text{i} \le 250 \\ \varphi_{ij} & \text{if } \ 251 \le \text{i} \le 500 \qquad \text{for } \ \text{j} = 1,2,\ldots,\text{J} \\ \zeta_{ij} & \text{otherwise}\end{cases} \tag{3.1}$$

$$Y_{ik} = \ln N\left(0,1\right) \ \text{for k} = 1,2,\ldots\text{K} \tag{3.2}$$

where $\eta_{ij} \sim \ln N\left(1.5,\sigma\right), \varnothing_{ij} \sim \ln N(-1.5,\sigma)$, and $\varsigma_{ij} \sim \mathit{ln}\ N\left(0,1\right)$

The choice of the sample sizes under condition 1 and 2, values of J and K, were chosen in order to cover a variety of situations that an experimenter may face when using real data and to be consistent with previous studies on microarray data. Sample sizes of (4,4) and (10,26) were chosen as in Kim et al. and Zhang's study where the latter is also the sample size of the Leukemia data from Baldi et al. [3]. The sample size (8,8) was also chosen since it is of same size as the apolipoprotein AI (Apo AI) dataset from Callow et al. [15]. For a thorough analysis covering more possibilities, sample sizes on a scale of 5 from 10 to 25 were also chosen for J and K. All of the sample sizes can be seen in Table 2. For the purpose of this study, the process of simulating a data set and running the methods under each setting was 500 times, while the previously mentioned studies of Zhang and Schwender et al. used 100 simulations for such comparisons.

Results and discussion

An advantage of simulating gene expression data is that the exact genes that are differentially expressed are known. After each method is performed on the simulated data sets, the total number of genes that were correctly identified as DE, true positives (TP), and the total number of genes that were incorrectly identified as DE, false positives (FP), were recorded. With the number of TP and FP known, then the type 1 error rate and the power were calculated to perform the comparison of methods. The null hypothesis for microarray analysis is that the ith gene under condition 1 is the same as under condition 2 i.e., it is not DE, versus the alternative where the ith gene under condition 1 is significantly different from the ith gene under condition 2 i.e., the ith gene is DE. The hypotheses are important to note in order to find the type 1 error rate, the probability of rejecting the null hypothesis given that it is in fact true, and the power, the probability of correctly rejecting a false null hypothesis. In terms of the microarray analysis done here the type 1 error rate reduces to the number of genes incorrectly identified as differentially expressed, FP, divided by the total number of equivalently expressed genes, 4500, and power reduces to the number

(*J*,*K*) and Skew	Power					P(type 1 error)				
	SAM	**sam-roc**	**FC**	**Med. FC**	**Rank Prod.**	**SAM**	**sam-roc**	**FC**	**Med. FC**	**Rank Prod.**
(4,4) L	0.0613	0.3986	0.4395	0.4445	0.4677	0.0022	0.0448	0.0427	0.0409	0.0809
M	0.0338	0.379	0.4538	0.4742	0.4528	0.0014	0.0433	0.044	0.0478	0.0791
H	0.018	0.3599	0.4638	0.479	0.4318	0.001	0.0413	0.0438	0.0474	0.0768
(8,8) L	0.3594	0.6436	0.7109	0.7235	0.6725	0.0116	0.0473	0.0472	0.047	0.0807
M	0.2415	0.584	0.6854	0.7072	0.6415	0.008	0.0452	0.0481	0.048	0.079
H	0.1224	0.5245	0.6468	0.6818	0.5965	0.0041	0.0417	0.0476	0.048	0.0772
(8,15) L	0.4194	0.7142	0.8144	0.8182	0.7677	0.0132	0.0492	0.0496	0.0492	0.0745
M	0.3806	0.65	0.7767	0.7907	0.728	0.0118	0.0477	0.05	0.0494	0.0728
H	0.3229	0.5843	0.7154	0.7521	0.6649	0.0098	0.0444	0.0497	0.049	0.0703
(8,20) L	0.4305	0.7243	0.8409	0.8579	0.8091	0.0134	0.0499	0.0472	0.0481	0.0716
M	0.4137	0.6596	0.8005	0.8225	0.7616	0.0125	0.0483	0.0475	0.0482	0.0691
H	0.3864	0.5935	0.7327	0.782	0.6932	0.0114	0.0462	0.0473	0.0485	0.0667
(8,25) L	0.4369	0.7208	0.8638	0.873	0.8383	0.0135	0.0498	0.0496	0.0489	0.0687
M	0.4341	0.6653	0.8248	0.8408	0.7874	0.0131	0.0489	0.0496	0.049	0.0663
H	0.4236	0.5999	0.7543	0.7961	0.7116	0.0126	0.0471	0.0498	0.0491	0.0637
(12,8) L	0.5427	0.7418	0.7979	0.8108	0.7452	0.0174	0.0481	0.049	0.0493	0.0813
M	0.397	0.6644	0.7409	0.7845	0.708	0.0126	0.0456	0.041	0.0485	0.0795
H	0.2371	0.5836	0.6944	0.7531	0.6531	0.0074	0.0427	0.0475	0.0484	0.0774
(10,15) L	0.5984	0.7959	0.8588	0.8601	0.8132	0.0187	0.0488	0.049	0.048	0.0773
M	0.4731	0.7204	0.8154	0.8284	0.7703	0.0144	0.0473	0.0489	0.0469	0.0754
H	0.3608	0.632	0.7427	0.7928	0.7028	0.0106	0.0443	0.0488	0.047	0.0732
(10,20) L	0.5895	0.8107	0.8905	0.8973	0.8559	0.0181	0.0496	0.0499	0.0468	0.0744
M	0.488	0.744	0.8498	0.8713	0.8099	0.0146	0.0482	0.0497	0.0491	0.0722
H	0.4168	0.6474	0.7674	0.829	0.7354	0.0121	0.0454	0.0498	0.0487	0.0699
(10,26) L	0.5567	0.8149	0.9068	0.9149	0.886	0.0168	0.0499	0.0486	0.0486	0.0722
M	0.4855	0.7487	0.8648	0.8861	0.8378	0.0145	0.0491	0.0484	0.0485	0.0699
H	0.4476	0.6574	0.7806	0.8432	0.756	0.0132	0.047	0.0488	0.0489	0.0673
(15,15) L	0.8165	0.8931	0.9176	0.9106	0.8735	0.0261	0.0489	0.0485	0.0482	0.0821
M	0.6801	0.8087	0.8747	0.8889	0.8318	0.021	0.047	0.0485	0.0483	0.0805
H	0.4591	0.6829	0.7765	0.8544	0.7588	0.0136	0.0434	0.0483	0.048	0.0789
(15,20) L	0.8581	0.916	0.9461	0.9471	0.9169	0.0267	0.0485	0.0496	0.0496	0.0794
M	0.7353	0.8465	0.9063	0.9261	0.8758	0.0226	0.0482	0.0497	0.0496	0.0778
H	0.5245	0.7177	0.8052	0.8913	0.7973	0.0153	0.0449	0.0495	0.0496	0.0762
(15,25) L	0.8775	0.9242	0.9596	0.9585	0.9433	0.0272	0.0492	0.0481	0.0485	0.0777
M	0.7625	0.8622	0.9219	0.939	0.9039	0.023	0.0481	0.0482	0.0485	0.0755
H	0.5577	0.7374	0.8211	0.9036	0.8243	0.0162	0.0462	0.0485	0.0487	0.0738
(20,20) L	0.9223	0.9537	0.9692	0.9735	0.9423	0.0289	0.0485	0.0495	0.0473	0.0829
M	0.8192	0.8876	0.9327	0.9592	0.9073	0.025	0.0475	0.0495	0.0473	0.0815
H	0.5906	0.7447	0.8247	0.935	0.8312	0.0172	0.0445	0.0497	0.0471	0.08
(20,25) L	0.9416	0.9634	0.9796	0.9827	0.9656	0.0295	0.0483	0.05	0.0493	0.0812
M	0.8542	0.9098	0.9491	0.9715	0.9346	0.026	0.0478	0.0499	0.0495	0.0797
H	0.6364	0.7702	0.8422	0.9488	0.8602	0.0184	0.0457	0.0497	0.0495	0.078
(25,25) L	0.9657	0.9784	0.9882	0.989	0.977	0.0303	0.0486	0.05	0.0488	0.0844
M	0.8903	0.9319	0.9624	0.981	0.9507	0.0268	0.0476	0.05	0.0488	0.0829
H	0.6616	0.7857	0.855	0.9644	0.8841	0.0187	0.0453	0.0499	0.0487	0.0815

Table 1: Power and P(type 1 error) for simulations under lognormal distribution.

of correctly identified differentially expressed genes, TP, divided by the total number of actual differentially expressed genes, 500.

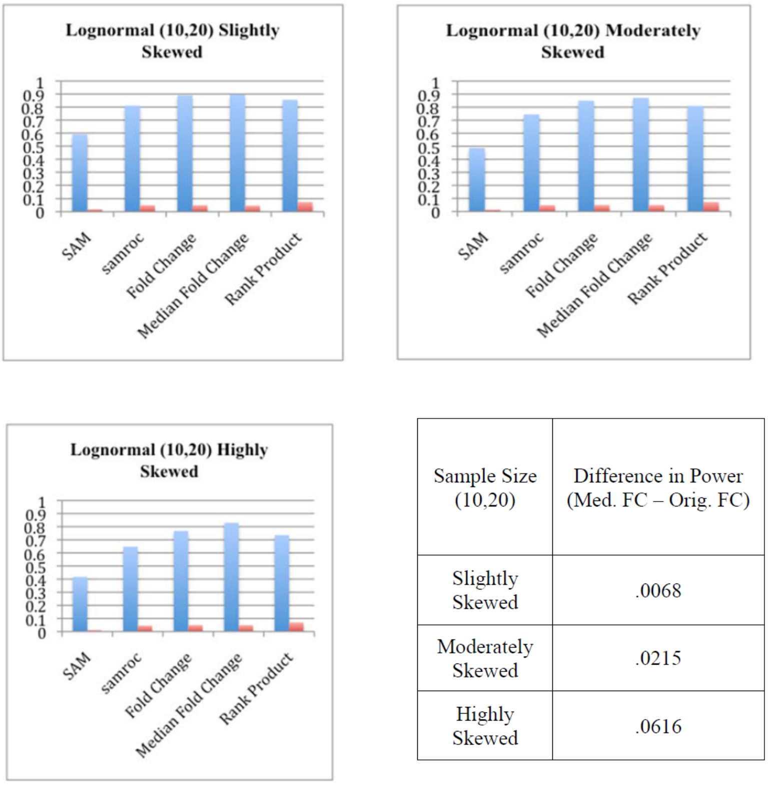
$$P(\text{type I Error}) = \frac{P(\text{reject null}\text{Ç}\text{null is true})}{P(\text{null is true})} = \frac{\text{FP}/5000}{4500/5000} = \frac{\text{FP}}{4500} \tag{3.3}$$

$$\text{Power} = \frac{P(\text{reject null}\text{Ç}\text{null is false})}{P(\text{null is false})} = \frac{\text{TP}/5000}{500/5000} = \frac{\text{TP}}{500} \tag{3.4}$$

The simulations carried out under the lognormal distribution revealed settings where the SAM method turns out to be the weakest of the methods. SAM worked rather poorly for all sample size combinations where at least one of the conditions had sample size less than 15. For settings where both conditions had 15 or more samples, SAM worked decently with a power most of the time above 0.70 except for few situations where the data was moderately skewed and in all cases that were highly skewed. In highly skewed settings, SAM was rather poor. The samroc method followed similar trends as SAM, however, samroc was more robust in respect to sample size. The performance of samroc was much better than SAM under settings where both conditions had sample sizes of 10 or higher. The values of power and type 1 error rate for each setting under a lognormal distribution are given in Table 1. Levels of skewness are indicated by L=slightly skewed, M=moderately skewed, and H=highly skewed (Table 1).

As Table 1 shows, the fold change method and the modified fold change method using median were consistently the top two methods across all sample sizes and all skewness settings for the lognormal data. The modified version of fold change with median worked better than the original fold change for all of the simulated sample sizes, obtaining higher levels of power while maintaining a type 1 error rate of 0.05 or smaller. It can also be seen in Table 1 that as the level of skewness rises, the modified version of the fold change method with median further improves over the original fold change. For each sample size simulated, as skewness increases, the difference in power between the original fold change and median fold change increases, with the latter having the higher power. This relationship is illustrated in Figure 1. The improvement in the fold change method was anticipated because the modified version replaced the mean with the median and for the lognormal data, which is a skewed distribution, the median is a more accurate measurement of the central tendency as Manikandan [14] stated.

Even though the median fold change method constantly had the better power as the sample size increased, it is evident that when there are at least 15 samples of each condition and the skewness is not too heavy, all the methods work very similarly, producing about the same power and type 1 error rate. The similar performance between methods toward the higher number of sample sizes leaves the decision of which method to use for analysis of microarray data to the researcher depending on which assumptions best match the data and the method of choice. SAM, samroc, and fold change all have the assumption that



Sample Size (10,20)	Difference in Power (Med. FC – Orig. FC)
Slightly Skewed	.0068
Moderately Skewed	.0215
Highly Skewed	.0616

Figure 1: Power and P(type 1 error) under lognormal distribution for sample size (10,20) for different levels of skewness. The blue columns correspond to the power and the red columns correspond to the P(type 1 error).

the genes share equal variance while rank product assumption is more relaxed allowing the variance to be about equal. The cutoff values for fold change and median fold change were chosen in order to obtain a type 1 error rate of no more than 0.05 and are given in Table 2.

Application

To illustrate the findings of this paper, a real data set, Apo AI data from Callow et al. [15] are analyzed in this section. The Apo AI dataset consists of 5548 genes and 16 samples. Out of the 16 samples, 8 were from control mice and the other 8 samples were from mice with the Apo AI gene knocked out. The 8 mice that had the Apo AI gene knocked out will have a very low high-density lipoprotein cholesterol level and the delivery of the cholesterol to the liver will be affected [15]. The data were preprocessed in the similar way as the leukemia data (4.1), as was done by Kim et al. The difficulty when attempting to analyze this dataset is that there has not been reference genes adopted as biologically significant from previous studies as there was with the leukemia data.

Table 3 shows the number of genes that were commonly found between each pair of the five methods. The idea expressed in Jeffery et al. [8] that only a very low percentage of genes will be found significant between multiple methods is supported by the results. The conflicting result between methods is one of the drawbacks of microarray analysis. There is a large inconsistency between the different methods to identify which genes are identified as significantly different between two groups [16] (Table 3).

Conclusion

A comparison of the performance of popular testing procedures for identifying differentially expressed genes from microarray data such as SAM, samroc, fold change and rank product was conducted. On the basis of the assumption that microarray data are related to the lognormal distribution from Hoyle et al. [13] and the familiar idea that the median is a better measurement of central tendency than the mean when describing skewed data as expressed in Manikandan [14], modifications were attempted on fold change, replacing the mean gene expression values with the median. It has been observed from Simulation results, fold change and the modified median fold change were consistently the top performing methods for lognormal data.

An analysis on a real microarray datasets was also performed to evaluate how the methods and the proposed modification would perform in a real situation. While the analysis on the Apo AI dataset showed that the median fold change method was an improvement to the original fold change, it also gave a nice visualization of how the different methods are inconsistent with each other when identifying differentially expressed genes. Hope findings of the paper will be useful for the practitioners in the area of health sciences and related area.

Sample Size	Fold Change	Median Fold Change
(4,4)	5.00	4.75
(8,8)	3.23	3.18
(8,15)	2.81	2.79
(8,20)	2.72	2.65
(8,25)	2.62	2.58
(10,10)	2.88	2.93
(10,15)	2.65	2.65
(10,20)	2.52	2.49
(10,26)	2.46	2.42
(15,15)	2.42	2.44
(15,20)	2.29	2.28
(15,25)	2.22	2.22
(20,20)	2.16	2.14
(20,25)	2.08	2.06
(25,25)	2.00	2.00

Table 2: Cutoff values for fold change and median fold change with at most 0.05 P(type 1 error) under lognormal distribution.

MethodS	SAM				
samroc	42	**samroc**			
Fold Change	0	25	**Fold Change**		
Median Fold Change	0	35	39	**Median Fold Change**	
Rank Product	33	182	1	3	**Rank Product**

Table 3: Number of common identified significant genes in the Apo AI dataset.

Acknowledgements

Authors are dedicating this paper to those who are suffering from cancer, all over the world.

References

1. Chu G, Narasimhan B, Tibshirani R, Tusher V (2002) SAM Significance Analysis of Microarrays-User guide and technical document [Online].
2. Majtán T, Bukovská G, Timko J (2004) DNA microarrays-techniques and applications in microbial systems. Folia Microbiol (Praha) 49: 635-664.
3. Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17: 509-519.
4. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Emperical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96: 1151-1160.
5. Broberg P (2003) Ranking genes with respect to differential expression. Genome Biol 4: 1-9.
6. Pan W (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics 19: 1333-1340.
7. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573: 83-92.
8. Jeffery IB, Higgins DG, Culhane AC (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 7: 359.
9. Kim SY, Lee JW, Sohn IS (2006) Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Stat Methods Med Res 15: 3-20.
10. Baker SC (2013) Next-generation sequencing vs. microarrays [Online]. Genetic Engineering and Biotechnology News.
11. McCarthy DJ, Smyth GK (2009) Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25: 765-771.
12. Koziol JA (2010) The rank product method with two samples. FEBS Lett 584: 4481-4484.
13. Hoyle DC, Rattray M, Jupp R, Brass A (2002) Making sense of microarray data distributions. Bioinformatics 18: 576-584.
14. Manikandan S (2011) Measures of central tendency: median and mode. J Pharmacol Pharmacother 2: 214-215.
15. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res 10: 2022-2029.
16. Hozo SP, Djulbegovic B, Hozo I (2005) Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 5: 13.