IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN:2278-3008, p-ISSN:2319-7676. Volume 13, Issue 3 Ver. III (May. – June. 2018), PP 86-94
www.iosrjournals.org

Microarray Data: A Powerful Method forIdentifying Differentially Expressed Genes

Md. Bipul Hossen[1*], Md. Javed Ali[1], Mst. Noorunnahar[2]
[1](Department of Statistics, Faculty of Science, Begum Rokeya University, Rangpur, Rangpur-5400)
[2](Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur 1706)
* Corresponding author- Md. Bipul Hossen(mbipu@brur.ac.bd)

Abstract : *Microarray technology observing thousands of gene expressions at once, has been the most popular research topics in recent decades. The new data promise to improve fundamental understanding of life on a molecular level and may prove very helpful in medical diagnosis, treatment and drug design. Identification of differentially expressed (DE) genes across tissue samples or experimental conditions for the analysis of microarray data is the greatest challenge nowadays. Several approaches have already been used to improve the identification of DE genes. In this study, the most popular methods such as Significance Analysis of Microarrays, two Samples Mean Test (t-test) and Wilcoxon Signed-Rank Sum test are applied to detect the DE genes in microarray cancer datasets. Our result shows a small number of common genes for the colon cancer, lung cancer and breast cancer analysis by using t-test, Wilcoxon Signed-Rank Sum test and Significance Analysis of Microarray, respectively. Among the analytical results, t- test provides the highest classification accuracy according to False Discovery Rate. Although all of these methods give similarly good results in the microarray data, t-test represents the best performance among them for real datasets. This study shows practical evaluation frameworks of checking powerful methods for identifying DE genes in microarray data.*
Keywords -*Microarray Technology, Identification of Differentially Expressed Genes, False Discovery Rate*

Date of Submission: 31-05-2018 Date of acceptance: 16-06-2018

I. Introduction

Microarray is a new and promising biotechnology which monitors expression levels for thousands of genes in cells simultaneously. Microarray experiments are carried with expressions of large number of genes and a little number of experimental samples. This unique data structure has been emerged as a completely new challenging area for the researchers because of its simultaneous high dimensionality and complexity with small sample size. Genes, which are differentially expressed (DE) within two experimental conditions, in a cell may need to be identified, for example between healthy patients and patients having cancer in the objective of illustration.Microarray analysis allows the researcher to recognize which genes are expressed differently between these two groups of patients, thus assisting to improve a treatment that aims these specific genes and to find out a more workable type of therapy.

Over the years many methods have been used to analyze the microarray data. This research paper describes statistical methods for the analysis of gene expression data from the given study of these methods can be classified into two types such as parametric and nonparametric methods. Examples of parametric methods are the t-test, Bayes t-test [1], an analysis of variance approach, and the B-statistic method [2]. The likelihood ratio principle in a nonparametric setting to construct two-sample test statistic [3]. The comparison of three model-free approaches, namely, nonparametric t-test, Wilcoxon (or Mann-Whitney) rank sum test, and a heuristic method based on high Pearson correlation and showed that these methods provide convenient and robust way [4]. A powerful simple method for finding differentially expressed genes and used stratification based tight clustering algorithm, principal component analysis, and information pooling [5]. The Bonferroni procedure, the Holm procedure [6], the Hochberg procedure [7], and the Westfall and Young procedure [8] address the multiple test problem by controlling the family-wise error rate, which is the probability that at least one false positive occurs over the collective tests. Many comparative strategies among different methods have been employed to find the most reliable one in identifying and detecting the highest proportion of the true DE genes [9].

In this study, the most popular methods such as Significance Analysis of Microarray (SAM), t- test and Wilcoxon Signed-Rank Sum test are used to identify the DE genes in microarray cancer datasets. t-test has been used as a parametric method, whereas Significance Analysis of Microarrays (SAM) and Wilcoxon Signed-Rank Sum test as the nonparametric methods.

Although, 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 significant genes are fairly huge in a data set, the detected number of significant genes by SAM is affected and the test becomes less powerful. t-test, inadequate sample size and following low power are commons problems. t-test can be used on very small samples, but it does not justify the use of very small samples unless larger sample sizes are impossible. The t-test should also not be used for multiple comparisons. On the other hand, Wilcoxon Signed-rank Sum test neither depends on the form of the parent distribution nor on its parameters. Any assumptions about the shape of the distribution are not required. For this case, this test is often applied as an alternative to the t-test. False discovery rate (FDR) is used for controlling the expected proportion of falsely rejected hypothesis. FDR is same as the family-wise error rate when all hypotheses are true, but is smaller otherwise [10].

During this study, 13, 163 and 149 common genes are found for the colon cancer, lung cancer and breast cancer analysis by using t-test, Wilcoxon Signed-Rank Sum test and SAM, respectively. t- test provides the highest classification accuracy according to FDR.

This paper is organized as follows. In section 2, the statistical techniques are briefly described. A simulation study under the different settings of sample size is performed on each of the methods in section 3.Section 4 discusses their application by analyzing a real data set. Finally, section 5 concludes the paper.

II. Methods

The above mentioned popular statistical methods for identifying DE genes in microarray datasets are reviewed in this section. The performance of the methods on data that follow a normal distribution. Let the i-th gene expression level of the j-th sample under condition 1 be represented byX_{ij} and the i-th gene expression level of the k-th sample under condition 2 be represented by Y_{ik}, where $j = 1, \dots, J$ and $k = 1, \dots, K$ represents replicates under condition 1 and 2, respectively. The gene number is represented by i, where $i = 1, \dots, n$. Colon cancer, lung cancer and breast cancer datasets investigated during this study consist of 22284, 22283 and 33297 genes, respectively.

2.1 Two Samples Mean Test (t-test)

The t-test is most commonly applied when test statistic follows a normal distribution with a known scaling term in the statistic. When the scaling term is unknown and is replaced by an estimated value based on the data, the test statistic follows a Student's t distribution. It can be used to determine if two sets of data are significantly different from each other. There are several versions of the two-sample t-test, depending on whether the sample size is large and whether it is reasonable to assume that gene expression levels have an equal variance under two conditions [11].

This test is used only when the two distributions have the same variance. To test whether the means are different, the t statistic can be calculated as follows

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S.\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad (1)$$

where, the pooled standard deviation is

$$S = \sqrt{\left(\frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}\right)} \quad (2)$$

Where $S_1^2 = \frac{1}{n_1-1}\sum(X_{ij} - \overline{X_1})$ and $S_2^2 = \frac{1}{n_2-1}\sum(X_{ij} - \overline{X_2})$

In the above formulae $n = n_1$and n_2 is number of participants of group one and group two respectively. The number of degrees of freedom for either group n-1, or the total sample size reduced by two (i.e.$n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.

2.2 Significance analysis of microarrays (SAM)

Significance analysis of microarrays (SAM) is a statistical technique, for determining whether changes in gene expression are statistically significant [12]. With the advent of DNA microarrays, it is now possible to measure the expression of thousands of genes in a single hybridization experiment. The data generated is considerable, and a method for sorting out what is significant and what isn't is essential. The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiments. SAM is a method for identifying genes on a microarray with statistically significant changes in expression, developed in the context of an actual biological experiment.

The response variable may be a group like untreated, treated (either unpaired or paired), a multiclass group (like breast, lung and colon cancer), a quantitative variable (like blood pressure) or a possibly censored survival time. SAM computes a statistic $d(i)$ for each gene $i$measuring the strength of the relationship between gene expression and response variable. It uses repeated permutations of the data to determine if the expression of any genes are significantly related to the response. The cutoff for the significance is determined by a tuning

parameter (delta), chosen by the user based on the false positive rate. One can also choose a fold change parameter to ensure that called genes change at least a pre-specified amount.
Test statistics of SAM is

$$d(i) = \frac{\bar{X}_I(i) - \bar{X}_U(i)}{s(i) + s_0} \quad (3)$$

$$s(i) = \sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{n_1 + n_2 - 2}(SS_I + SS_U)} \quad (4)$$

Where $\bar{X}_I(i)$ and $\bar{X}_U(i)$ are mean expressions of gene i in condition I or U, $s(i)$ is gene specific scatter ands_0 is a small positive constant calculated to minimize coefficient of variation.
By replacing (4) into (3), it can be obtained that

$$d(i) = \frac{\bar{X}_I(i) - \bar{X}_U(i)}{\sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{n_1 + n_2 - 2}(SS_I + SS_U)} + s_0} \quad (5)$$

2.3 Wilcoxon Signed-Rank Sum Test

The Wilcoxon Signed-Rank Sum test is a non-parametric statistical hypothesis test used while comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed.
Let us consider$x_{1,i}$ and $x_{2,i}$denote the measurements and N be the sample size, i.e., the number of pairs. Thus, there are a total of 2N data points. For pairs$i = 1, 2, 3, \ldots, N$. The other parameters are
H_0: difference between the pairs follows a symmetric distribution around zero
H_1: difference between the pairs does not follow a symmetric distribution around zero.
The test procedure follows

- For $i = 1, 2, 3, \ldots \ldots N$,calculate $|x_{2,i} - x_{1,i}|$and$sgn(x_{2,i} - x_{1,i})$where sgn is the sign function.
- Exclude pairs with $|x_{2,i} - x_{1,i}| = 0$. Let $N_r$be the reduced sample size.
- Order the remaining $N_r$pairs from the smallest absolute difference to the largest absolute difference, $|x_{2,i} - x_{1,i}|$.
- Rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let R_i denote the rank.
- Calculate the test statistic W, W=The sum of the signed ranks= $\sum_{i=0}^{N_r}[\mathrm{sgn}(x_{2,i} - x_{1,i}).R_i]$.
- Under null hypothesis, W follows a specific distribution with number of simple expression. This distribution has an expected value of 0 and a variance of $\frac{N_r(N_r+1)(2N_r+1)}{6}$. It is compared to a critical value from a reference table.
- The two-sided test consists in rejecting H_0, if $|w| \geq W_{critical,\ N_r}$.
- As N_r increases, the sampling distribution of W converges to a normal distribution. Thus, For $N_r \geq 10$, a z-score can be calculated as$z = \frac{W}{\sigma_W}$, $\sigma_W = \sqrt{\left(\frac{N_r(N_r+1)(2N_r+1)}{6}\right)}$

If $|z| \geq z_{critical}$ then reject H_0 (two-sided test).
Alternatively, one-sided tests can be realized with either the exact or the approximate distribution. p-value can also be calculated.

Alternatively, one-sided tests can be realized with either the exact or the approximate distribution. p-value can also be calculated.
The t statistic [13] is the smaller of two sums of ranks of given sign. Low values of t are required for significance. t statistic is easier to calculate than W test and the test is equivalent to the above-described two-sided test (the distribution of the statistic under H_0 has to be adjusted). The details of this method can be found in [14-16].

III. Results and Discussion

To demonstrate the performance of different methods (SAM, Wilcoxon Signed-Rank Sum Test and t-test), both simulated and real microarray gene expression datasets as well as five R packages of other methods (such as samr, genefilter, Biobase, multtest and siggenes) are used. FDR are computed then for each method. All R packages are available in the comprehensive R archive network (cran) or bioconductor.

3.1 Data Generating Model

Suppose that Y_{jk} is the expression level of gene *j* in array k ($j = 1, \dots n$; $k = 1, \dots, k_1, k_1 + 1, \dots k_1 + k_2$) and the first k_1 and last k_2 arrays are obtained under two conditions, respectively. A general statistical model can be written as

$$Y_{jk} = a_j + b_j X_k + \varepsilon_{jk}$$

Where $X_k = 1$ for $1 \leq k \leq k_1$ and $X_k = 0$ for $k_1 + 1 \leq k \leq k_1 + k_2$, and ε_{jk} are random errors with mean 0. Hence $a_j + b_j$ and a_j are the mean expression levels of gene *j* under two conditions, respectively.

3.1.1 Simulation Study in Simulated Data

To investigate performance of the chosen analysis as applied to the observed data, it has generated microarray datasets by the model as displayed in Table 1. This dataset has three different levels corresponding two different groups. In Table 1, the number of columns are represented the individuals {Group-1 and Group-2} are 'healthy' and 'sick' patient group and the number rows are represented the genes. The numbers of first (1-10) rows are DE genes; second (11-20) rows are DE genes and third (21-200) rows are EE genes. To randomize the gene expressions among the individuals, it has randomly added Gaussian noise from $N(0, \sigma^2)$ with each expression of each gene.

Table 1. Matrix used to generate simulation study

Number of row	**Group-1 (sample size=22)**	**Group-2 (sample size=22)**	
1-10	+a	-a	$+ N(0, \sigma^2)$
11-20	-a	+a	$+ N(0, \sigma^2)$
21-200	a	a	$+ N(0, \sigma^2)$

Now we make a simulation study in generated a data following way,
Total Number of Generated Genes G=200
Number of Equally Expressed (EE) genes = 180 (Generated under H_0)
Number of Differentially Expressed (DE) genes =20 (Generated under H_1)
Number of samples in condition-1= 22
Number of samples in condition-2=23
Percentage of contaminated Genes =10% and
Percentage of contaminated samples in each gene 5%
To examine the performance of the existing methods, we added some randomly contaminated genes having outlying expressions in above generated model.
In this study, it uses t-test, Wilcoxon signed-rank sum test and SAM methods in frequently by using R language directory command and then find 17,19 and 17 DE genes at 3.0 delta values out of 200 genes respectively. All of these methods provide highest accuracy in simulation data.

3.2 Real Microarray Data

To evaluate the performance of the method in a comparison among the methods as mentioned earlier, it has used three microarray datasets. Colon Cancer dataset is the first dataset which consists of 21 samples with 22,284 genes. The second dataset is the Lung Cancer dataset, which contains 22,283 genes from 45 samples and the last dataset is the Breast Cancer dataset, which consists of 33,297 genes for 19 breast cancer samples.

3.2.1 Colon cancer Dataset

Colon cancer gene expression dataset [17] is used in this section. Fig. 1.1(a), Fig. 1.1(b) and Fig.1.1(c) represent the Q-Q plots for this dataset using Wilcoxon signed-rank sum test, t-test and SAM method, respectively. In this figure, the number of genes above the band in the upper right and below the band in the bottom left (green color) indicates the number of up regulated and down regulated DE genes, respectively. It shows 15, 58 and 437 genes at delta value 0.4 and provides 13 common differentially expressed genes those are detected by applying three methods in this dataset. Table 2 and Fig.1.2 show that FDR of the first three values of t-test are small. Although the last three values of Wilcoxon signed-rank sum test are small, the variation between the third and fourth values is large. Thus the t-test is the best identifying method for colon cancer data.

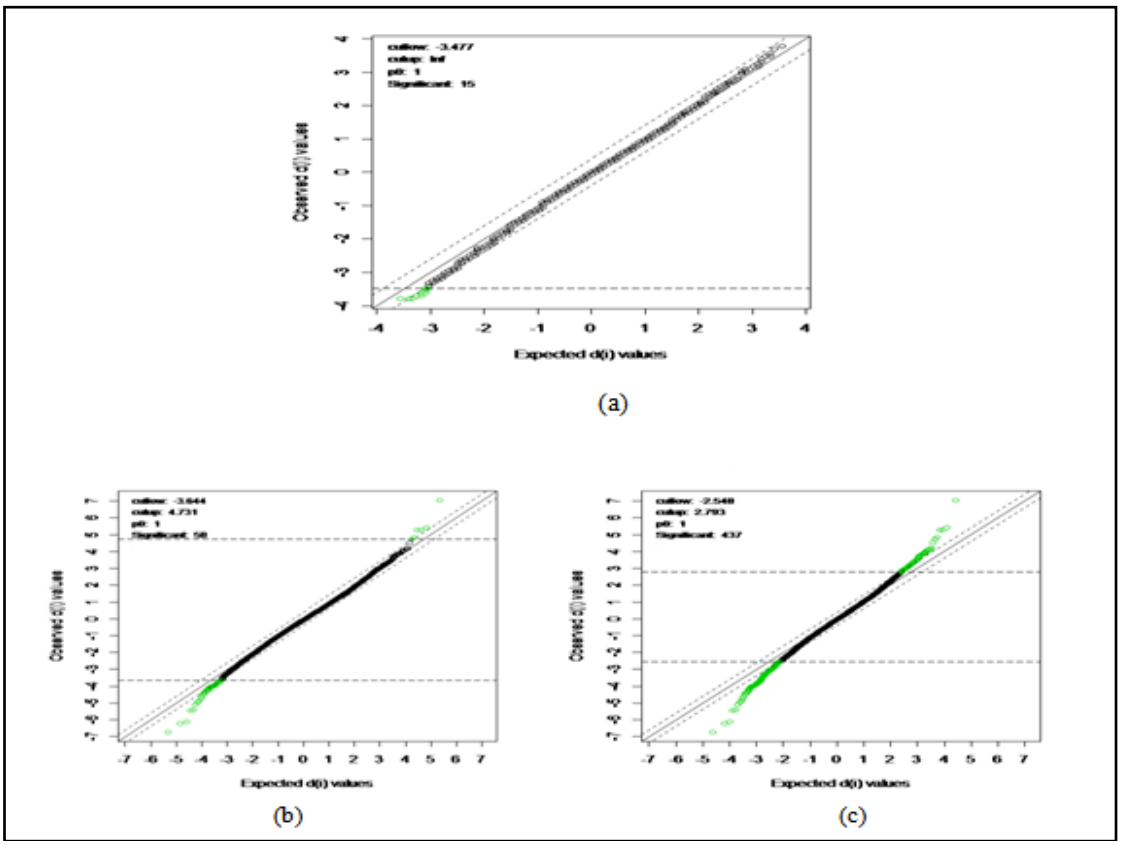

Figure 1.1. Performance evaluation using Q-Q plot for detection of DE genes using (a) Wilcoxon signed-rank sum test, (b) t-test and (c) SAM methods.

Table 2. False Discovery Rate of different methods.

Delta	Wilcoxon Signed-Rank Sum Test	t-test	Significance Analysis of Microarray
0.1	0.885	0.827	1
0.2	0.688	0.638	1
0.3	0.559	0.471	0.942
0.4	0.095	0.389	0.798
0.5	0.053	0.307	0.655
0.6	0	0.201	0.2006

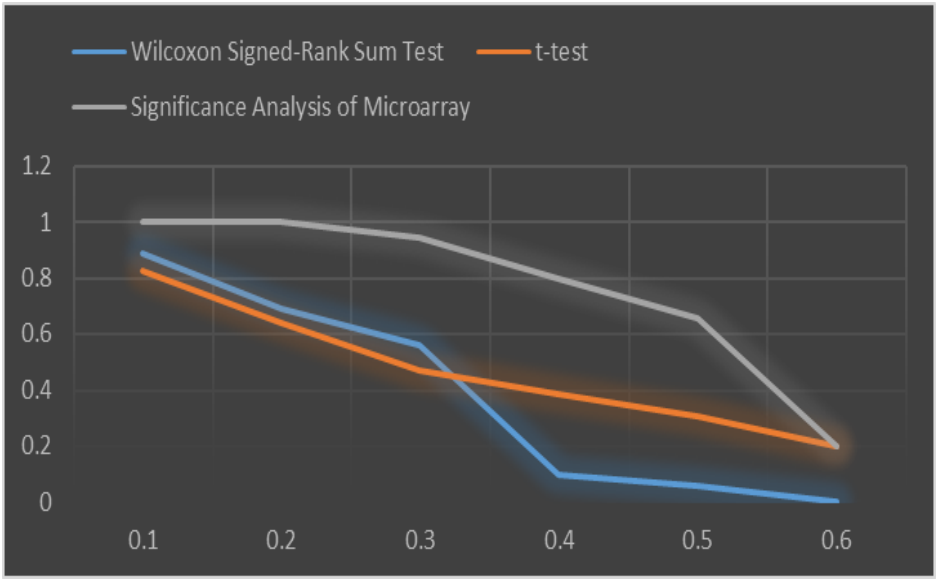

Figure 1.2. False Discovery Rate of different methods as a function of delta.

3.2.2 Lung Cancer Dataset

Lung cancer gene expression dataset [18]is also used in this study. DE genes have been obtained from Wilcoxon signed-rank sum test, t-test and SAM methods, which are 164, 1650 and 2016 at 1.7 delta value as presented in Fig. 2.1(a), Fig. 2.1(b) and Fig. 2.1(c) respectively. In this case, 163 are common genes, those are detected by all these methods, and some of the genes are responsible for cancer disease. As illustrated in both Table 3 and Fig. 2.2, FDRs of t-test and SAM are approximately same.

Table 3. False Discovery Rate of the investigated methods.

Delta	Wilcoxon Signed-Rank Sum Test	t-test	Significance Analysis of Microarray
0.1	0.542	0.456	0.461
0.3	0.430	0.359	0.361
0.5	0.299	0.259	0.256
0.7	0.198	0.0.171	0.165
0.9	0.109	0.103	0.096
1.1	0.051	0.0.058	0.052
1.3	0.020	0.030	0.026
1.5	0.005082	0.015	0.012
1.7	0.000567	0.007	0.005
1.9	0	0.003	0.002

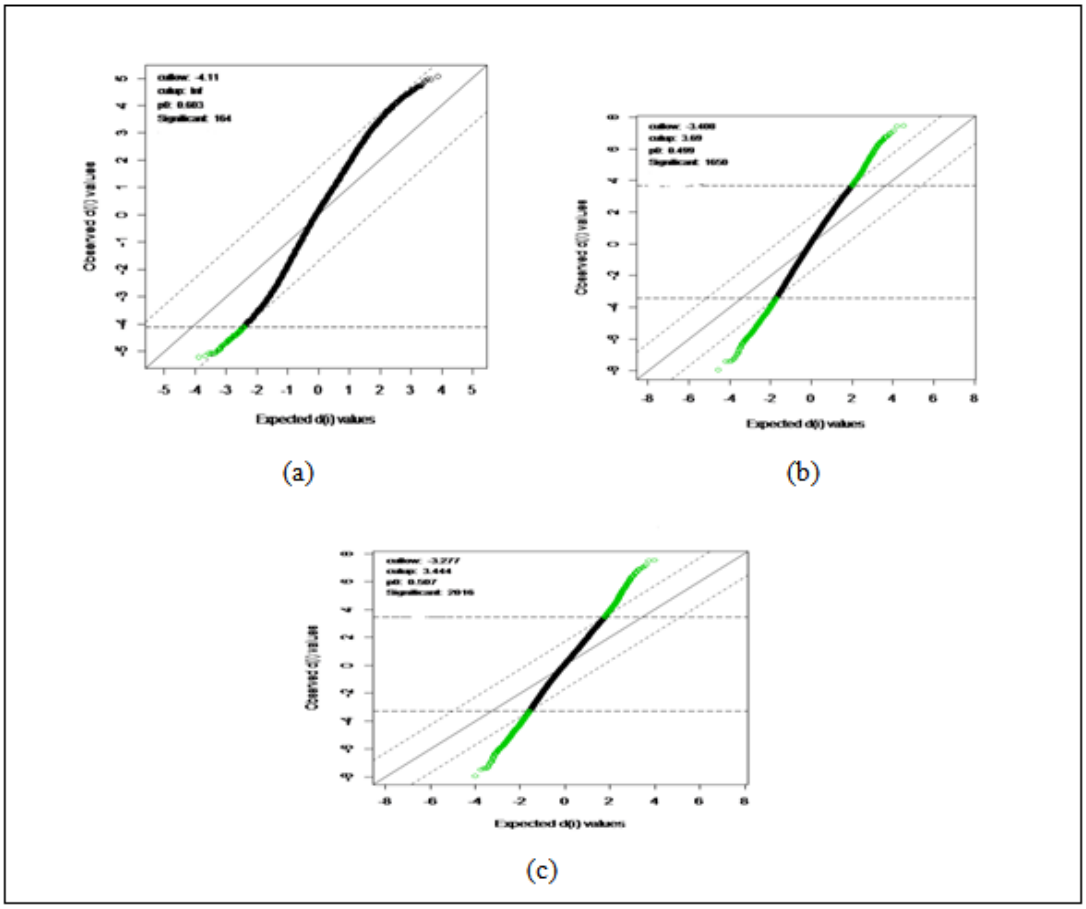

Figure 2.1. Performance evaluation using Q-Q plot for detection of DE genes using (a) Wilcoxon signed-rank sum test, (b) t-test and (c) SAM methods.

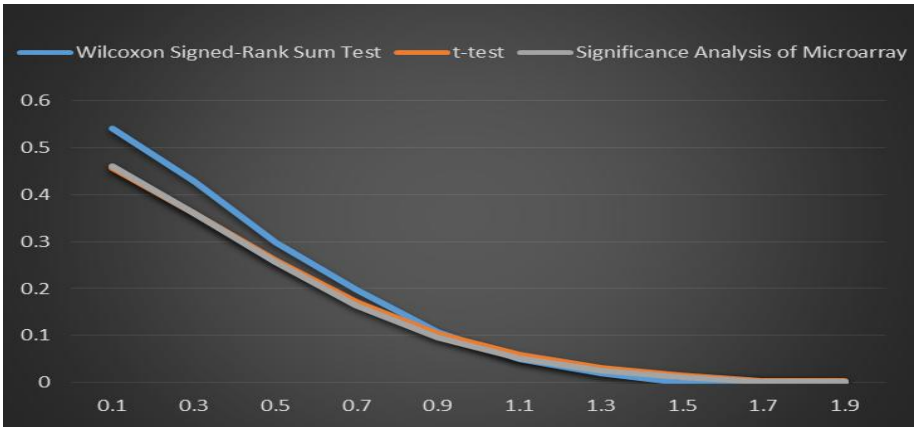

Figure 2.2. False Discovery Rate versus delta of the investigated methods.

3.2.3 Breast Cancer Dataset

Breast cancer dataset[19] is a more severe cancer and we found from this dataset, 151, 1022 and 1331 DE genes are detected by Wilcoxon signed-rank sum test, t-test and SAM at 0.5 delta value as shown in Fig. 3.1(a), Fig. 3.1(b) and Fig. 3.1(c) respectively. 149 genes that are commonly found between each pair of the three methods. t-test test provides lower FDR as presented in Table 4 and Fig. 3.2. Although SAM gives almost similar values as t-test, it can be easily concluded that t-test is the best for breast cancer data.

Table 4. False Discovery Rate of the methods.

Delta	Wilcoxon Signed-Rank Sum Test	t-test	Significance Analysis of Microarray
0.1	0.907	0.6773	0.7481
0.2	0.783	0.573	0.589
0.3	0.638	0.4593	0.4736
0.4	0.506	0.357	0.374
0.5	0.137	0.278	0.2902
0.6	0	0.2148	0.228

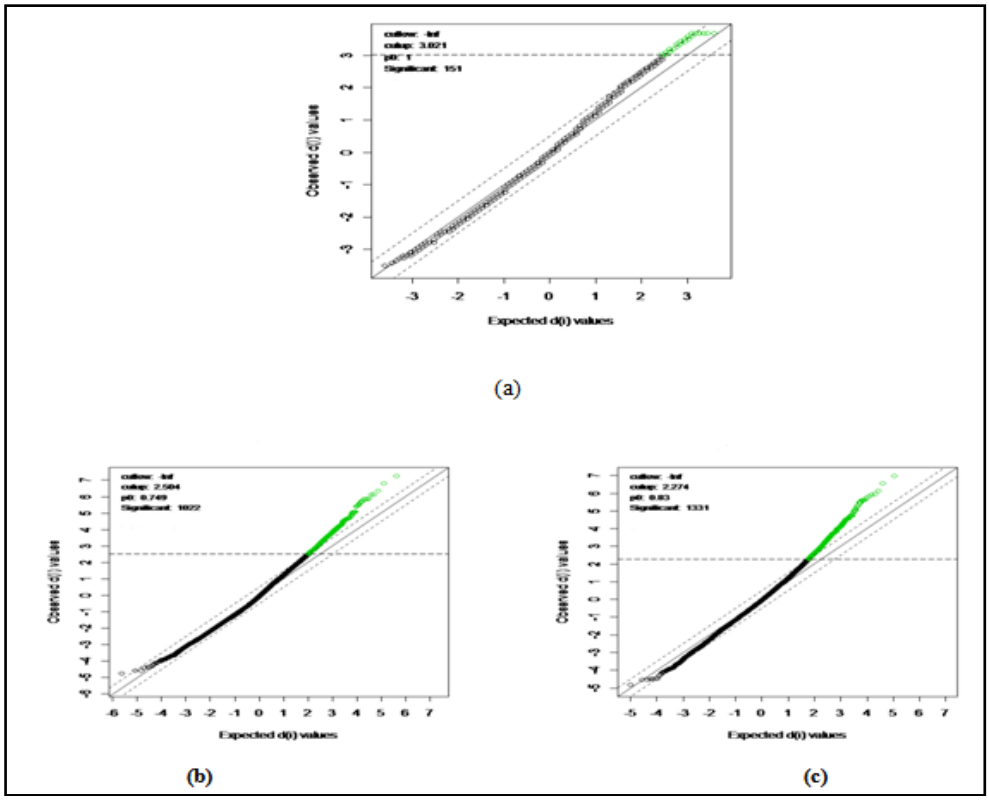

Figure 3.1. Performance evaluation using Q-Q plot for detection of DE genes using (a) Wilcoxon signed-rank sum test, (b) t-test and (c) SAM methods.

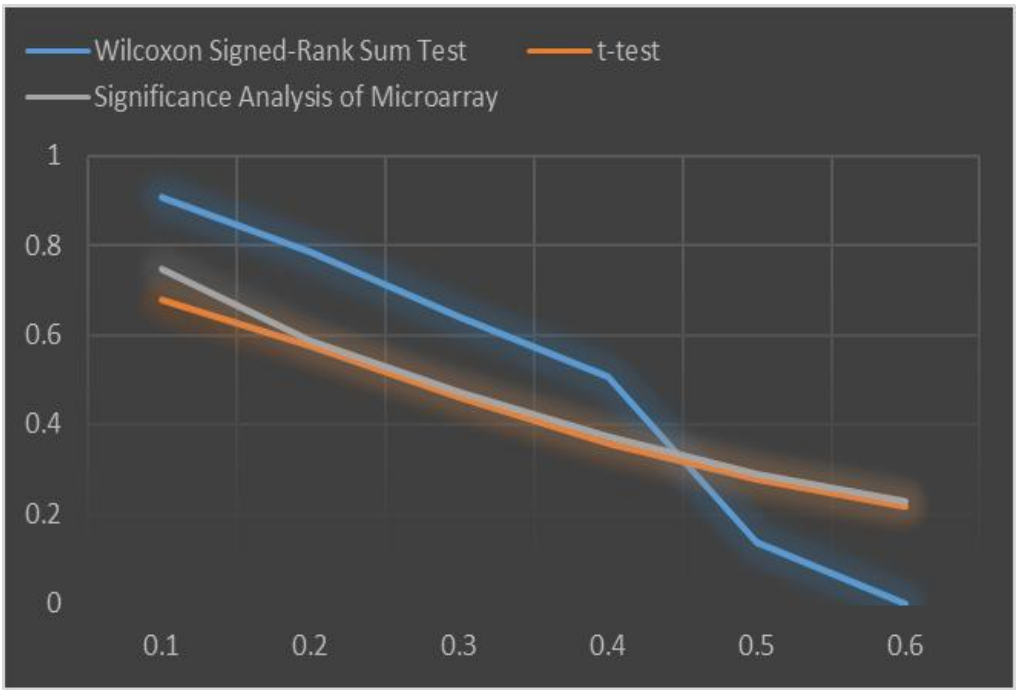

Figure 3.2.False Discovery Rate of the methods

IV. Conclusions

A comparison of the performance of popular methods such as SAM, t-test and Wilcoxon signed-rank sum test have been applied for identifying DE genes from microarray data. It has been observed from simulation results, all methods are consistently the best performing methods for microarray data. Three real microarray datasets are also performed to evaluate which identifying method is the best in a real situation. The analysis on the datasets has showed that t-test is the best method for identifying DE genes among the investigated methods.

Acknowledgements

The authors would like to thank the reviewers in advance for their valuable comments and suggestions.

References

[1]. Baldi, P., Long, A.D.; A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001; 17: 509-519.
[2]. Smyth, G.K.; Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 2004; 3: Article 3.
[3]. Bokka, S. and Mathur, S.; A nonparametric likelihood ratio test to identify differentially expressed genes from microarray data. Appl Bioinform 2006; 5(4): 267–276.
[4]. Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D and Altman, R.B.; Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002; 18(11): 1454–1461.
[5]. Qin, H., Feng, T., Harding, S.A., Tsai, C-J and Zhang, S.; An efficient method to identify differentially expressed genes in microarray experiments. Bioinformatics 2008; 24(14): 1583–1589.
[6]. Holm,S.; A simple sequentially rejective multiple test procedure, Scand. J. Stat.1979; 6: 54–70.
[7]. Hochberg,Y.; A sharper Bonferroni procedure for multiple tests of significance, Biometrika 1988; 75: 800–803.
[8]. Westfall,P., Young,S.; Resampling-Based Multiple Testing, Wiley, New York, 1993.
[9]. Md. Bipul Hossen, Md. Siraj-Ud-Doulah. Identification of Robust Clustering Methods in Gene Expression Data Analysis. Current Bioinformatics 2017; 12(6):558-562
[10]. Benjamini, Y. and Hochberg, Y.; Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc.B 1995; 57:289–300.
[11]. Devore, J. And Peck, R.; Statistics: The exploration and analysis of data. 3rd edition, 1997.
[12]. Tusher, V.G., Tibshirani, R., and Chu, G.; Significance Analysis of Microarrays Applied to the Ionizing Radiation Response, Proceeding National Academy of Sciences USA 2001; 98(9):5116-5121.
[13]. Siegel, Sidney; Non-parametric statistics for the behavioral science. New York: McGraw-Hill,1956.
[14]. Jung, K., Quast, K., Gannoun, A. and Urfer, W.; A renewed approach to the nonparametric analysis of replicated microarray experiments. Biometrical Journal 2005; 48: 245-254.
[15]. Troyanskaya, O.G., Barber, M.E., Brown, P.O., Botstein, D., Altman, R.B; Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002; 18:1454–1461.
[16]. Wu,T.D; Analysis gene expression data from DNA microarrays to identify candidate genes. Journal of Pathology 2001; 195:53–65.

[17]. Koinuma, K., Yamashita, Y., Liu, W., Hatanaka, H.; Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. Oncogene 2006; 25(1):139-146.
[18]. Coldren, C.D., Helfrich, B.A., Witta, S.E., Sugita, M.; Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. Mol Cancer Res 2006; 4(8):521-528.
[19]. Yang, L., Wu, X., Wang, Y., Zhang, K.; FZD7 has a critical role in cell proliferation in triple negative breast cancer. Oncogene 2011 30(43):4437-46.

IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS) is UGC approved Journal with Sl. No. 5012, Journal no. 49063.

Md. Bipul Hossen "Microarray Data: A Powerful Method for Identifying Differentially Expressed Genes." IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS) 13.3 (2018): 86-94.