J. Bio-Sci. 25: 9-16, 2017 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index

A COMPARATIVE STUDY OF BIOMARKER GENE SELECTION METHODS IN PRESENCE OF OUTLIERS

M Shahjaman[1,2*], N Kumar[1,3], AA Begum[1], SMS Islam[4] and MNH Mollah[1*]

[1]Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh; [2]Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh; [3]Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh; [4]Institutitute of Biological Sciences, University of Rajshahi, Bangladesh

Abstract

The main purpose of gene expression data analysis is to identify the biomarker genes by comparing the gene expression levels between two different groups or conditions. There are several methods to select biomarker genes and many comparative studies have been performed to select the appropriate method. However, they did not consider the problems of outliers in their data sets though it is very essential to select the method from robustness point of view due to outliers may occur in the different steps of the gene expression data generating process. In this paper, it is evaluated the performance among five popular statistical biomarker gene selection methods *viz*. T-test, SAM, LIMMA, KW and FCROS using both simulated and real gene expression data sets in absence and presence of outliers. In the simulated data analysis, it was demonstrated the performance of these methods in terms of different performance measures such as TPR, TNR, FPR, FNR and AUC and based on these measures, it was found that in absence of outliers, for both small-and-large sample cases all the methods perform almost similar. Whereas, in presence of outliers, for small-sample case only the FCROS method perform well than other methods. From a real colon cancer data analysis, it was elucidated that FCROS method identified additional 59 genes that were not detected by the other methods and most of them belongs to the different cancer related pathways.

Key words: Biomarker genes, DE genes, FCROS, outliers, robustness

Introduction

Microarrays gene expression data analysis is a promising area of bioinformatics. It allows simultaneous measurement of the expression levels of thousands of genes. To generate the gene expression data there are many ways, the most popular way is the so called DNA microarray technology. Microarray gene expression data can be viewed as a matrix of $m \times n$ dimension, organized by m genes versus n samples (patients) after completing several steps with the help of biological technology and statistical learning. In general m may have 10 - 100 thousands of genes and n can have 3 - 30 samples. This unique data structure has discovered as a completely new area of research for both statisticians and biologists. At the same time it provides a challenge to researchers because of high dimensionality and its complexity (large m and small n problem). The main purpose of the microarray experiments is to select the biomarker genes by comparing levels of gene expression between two different groups/conditions (Kaissi et al. 2013). In general, one group is known as reference and other is experiment. There are several methods available in the literature in this regard and many studies have been performed to select the appropriate method among these methods. For

*Author for correspondence: shahjaman_brur@yahoo.com

example Cui et al. (2003) performed a comparative study among the popular gene selection methods such as T-test, SAM-significance analysis of microarray (Tusher et al. 2001), LIMMA (Efron et al. 2001), F-test (Kerr et al. 2000) and Kruskal Wallis (KW) (Kruskal et al.1952). Dembele and Kastner (2014) developed a new approach called fold change rank ordering statistic (FCROS) based on the FC ranks between two experimental groups. However, most of the approaches discussed above are not robust against outliers. Also, most of the comparative studies did not consider the problems of outliers in their datasets. Outliers are often occur in the gene expression data due to several steps involved in the data generating process from hybridization to image analysis (Shahjaman et al. 2017). Thus in presence of outliers, the results of downstream analysis using the popular gene selection methods might be changed. Mollah et al. (2015) reported that the assumption of normality of many gene selection methods do not hold for some existing microarray datasets in presence of outliers. Though, KW and FCROS are robust against outliers for large sample case but they are sensitive to outliers for small-sample case. Furthermore, it is very difficult to identify few biomarker genes from the high-dimensional outlying gene expression dataset and there is no precise guideline about the selection of methods. Therefore, in this paper, it was evaluated the performance among the popular biomarker gene selection statistical methods based on microarray gene expression data set in absence and presence of outliers to select the proper method.

Materials and Methods

In order to assess the performance of all the gene selection methods, we consider the different measures such as true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR) and area under the Receiver Operating Characteristics curves (AUC). In a two groups prediction problem such as DE (differentially expressed) or EE (equally expressed), the outcomes are divided into four categories: (i) truly DE that are reported as DE (True Positives: TP), (ii) truly EE that are reported as DE (False Positive: FP), (iii) truly DE that are reported as EE (False Negative: FN) and (iv) truly EE that are reported as EE (True Negatives: TN). Then the formulas of all the measures are as follows:

TPR = TP/TP +FN, TNR= TN/TN +FP, FPR= FP/TN +FP, FNR= FN/FN +TP.

Three R packages were used as (i) *limma,* which was proposed by Smyth et al. (2003), (ii) *samr,* which was developed by Tibshirani et al. (2013) and (iii) *ROCR* to obtain an area under a ROC curve (AUC) by Sing et al. (2005). A method will be called good performer that produces the higher values of TPR, TNR and AUC and lower values of FPR and FNR. We declared a gene as DE gene with adjusted *p*-value <0.05. *P*-values are adjusted by the Benjamini and Hochberg (1995) multiple testing correction method.

Simulated dataset

We applied five biomarker gene selection methods in the simulated dataset. The simulated dataset was generated using the following one-way ANOVA model developed by Kerr et al. (2000):

$$x_{jk} = \mu_j + \epsilon_{jk}; (j = 1,2; k = 1,2, \dots, n_j) \quad (1)$$

where, x_{jk} is the *k*th observed expression of a gene in the *j*th condition, μ_j is the mean of all expressions of a gene in the *j*th condition and ϵ_{jk} is the random error term that follows$N(0, \sigma^2)$.The outlying dataset was generated by multiplying a constant (say, 5) with the mean of equation (1).

Colon cancer real dataset

To investigate the performance of the five methods as early mentioned, in the real microarray gene expression data was used colon cancer data set was used which consists of 22 normal and 40 tumor samples. Alon et al. (1999) was used this dataset in their study. This dataset contains 2000 genes.

Results and Discussion

To investigate the performance of the five biomarker gene selection methods 100 data sets were generated from one-way ANOVA model using equation (1) for both small ($n_1 = n_2 = 3$) and large-sample ($n_1 = n_2 = 15$) cases with two groups.

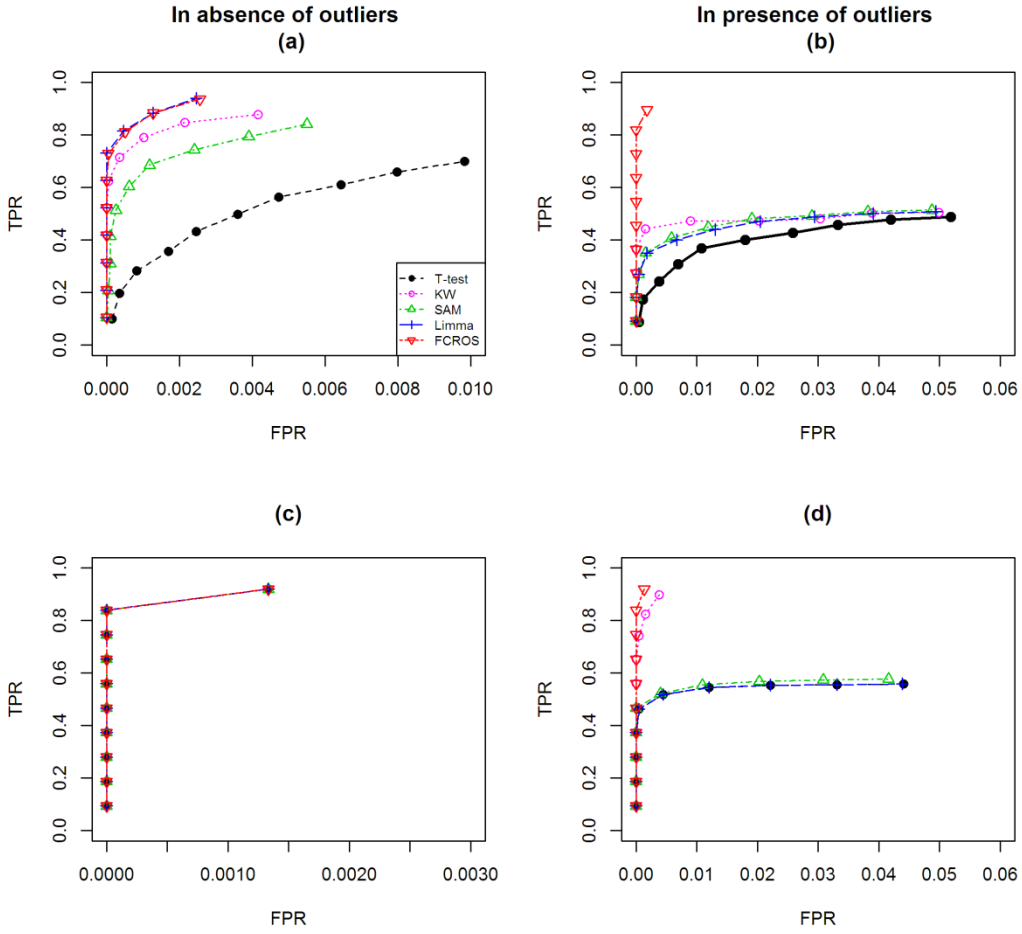

Fig. 1. ROC curve produced by different methods based on simulated gene expression data set. For small-sample case ($n_1 = n_2 = 3$): (a) in absence of outliers, (b) in presence of outliers. For large-sample case ($n_1 = n_2 = 15$): (c) in absence of outliers and (d) in presence of outliers.

In this case, 10000 genes were generated with ($n_1 + n_2$) samples for each of the dataset. The number of DE gene is set to 300 and the rest of the 9700 genes are considered as the EE genes. The arbitrary values $(\mu_1, \mu_2) \in c(3, 5)$ and σ^2= 0.1 were set. Each dataset for each case were represented the gene expression profiles of 10,000 genes and ($n_1 + n_2$) samples. Fig.1a and c represents the ROC curve produced by different methods in absence of outliers for both small-sample size ($n_1 = n_2 = 3$) and large-sample size ($n_1 = n_2 =15$), respectively. It is clear from these figures that in absence of outliers all the methods produce similar results, except T-test for small-sample case. But in presence of 5% outliers, for small-sample case ($n_1 = n_2 =3$), the performance of all the methods has significantly deteriorated.

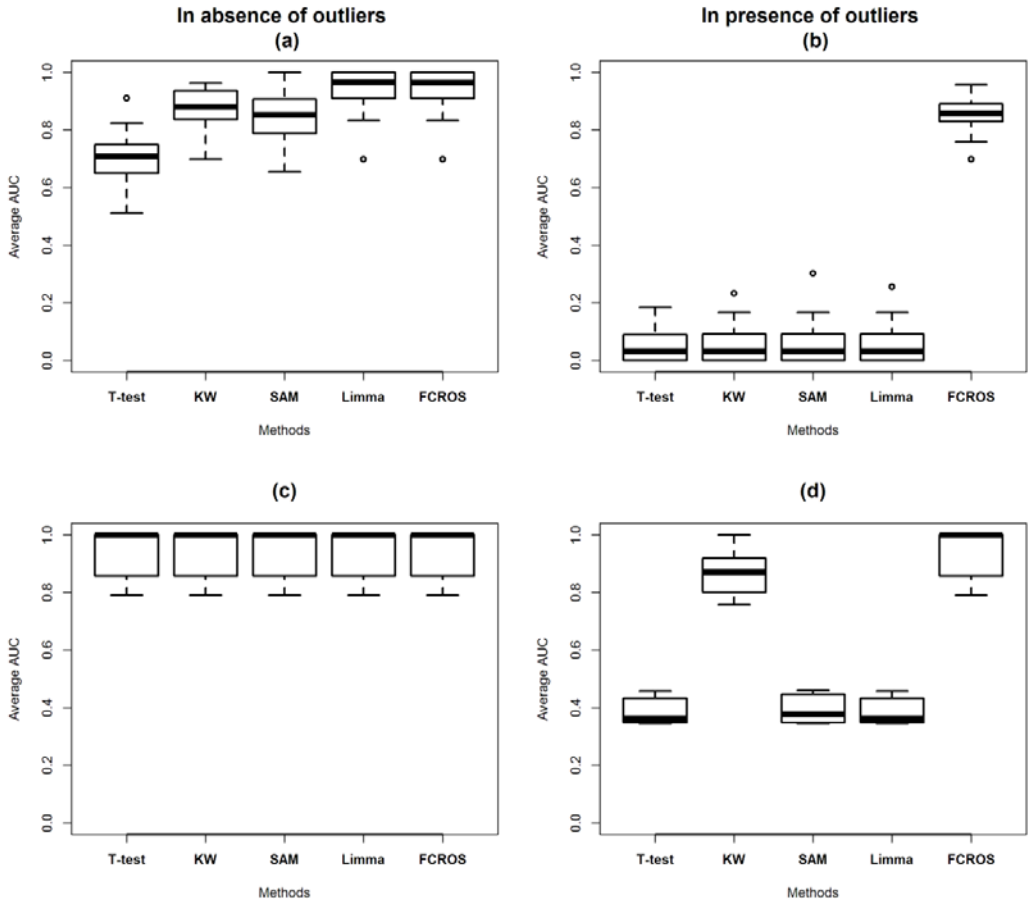

Fig. 2. Box plot of AUC values produced by different methods based on simulated gene expression data. For small-sample case ($n_1 = n_2 = 3$): (a) in absence of outliers, (b) in presence of outliers. For large-sample case ($n_1 = n_2 =15$): (c) in absence of outliers and (d) in presence of outliers.

In this case only FCROS method performed well (Fig.1b). Whereas, for large-sample case ($n_1 = n_2 =15$) (Fig. 1d) with 5% outliers, two methods FCROS and KW performed well. FCROS performed slightly better than KW in this case. Fig. 2a and c shows the box-plots of AUC values for each of the methods based on 100 simulated datasets generated using equation (1) in absence of outliers for small-and-large sample cases,

respectively. It is evident from Fig. 2a that, in absence of outliers for small sample case LIMMA, FCROS, SAM and KW performed better than T-test. Whereas, in case of Fig. 2c, for large sample case every methods performed similar. Fig. 2b and d represents the box-plots of AUC values based on 100 simulated datasets in presence of 5% outliers for both small-and large-sample cases respectively. It is apparent from Fig. 2b that, for small-sample case, the performance of all the methods has declined significantly except FCROS. However, for large-sample case in presence of outliers (Fig. 2d) two methods FCROS and KW performed well than the other three methods (T-test, SAM and LIMMA). Table 1 summarizes the average values of the different performance measures TPR, FPR, TNR, FNR and AUC based on 100 simulated datasets for both small-and-large sample cases in absence and in presence of outliers, respectively.

Table 1. Performance evaluation based on simulated dataset with 2 groups.

For small sample size ($n_1 = n_2 = 3$)										
	In absence of outliers					In presence of outliers				
Methods	TPR	FPR	TNR	FNR	AUC	TPR	FPR	TNR	FNR	AUC
T-test	0.748	0.008	0.992	0.252	0.746	0.343	0.02	0.98	0.657	0.342
SAM	0.892	0.004	0.996	0.108	0.891	0.374	0.019	0.981	0.626	0.373
LIMMA	0.947	0.002	0.998	0.053	0.947	0.364	0.019	0.981	0.636	0.364
KW	0.860	0.005	0.995	0.140	0.860	0.351	0.020	0.98	0.649	0.351
FCROS	0.943	0.002	0.998	0.057	0.943	0.915	0.003	0.997	0.085	0.915
For large sample size ($n_1 = n_2 = 15$)										
	In absence of outliers					In presence of outliers				
Methods	TPR	FPR	TNR	FNR	AUC	TPR	FPR	TNR	FNR	AUC
T-test	0.939	0.003	0.997	0.061	0.939	0.379	0.019	0.981	0.621	0.379
SAM	0.939	0.003	0.997	0.061	0.939	0.395	0.019	0.981	0.605	0.395
LIMMA	0.939	0.003	0.997	0.061	0.939	0.382	0.019	0.981	0.618	0.382
KW	0.939	0.003	0.997	0.061	0.939	0.855	0.006	0.994	0.145	0.854
FCROS	0.939	0.003	0.997	0.061	0.939	0.939	0.003	0.997	0.061	0.939

It is noticeable from this table that, for small-sample case in absence of outliers every method performed well except T-test while in presence of outliers, only FCROS performed well than the other four methods (T-test, KW, SAM and LIMMA) methods. For example, FCROS produces AUC >0.90 than T-test, KW, SAM and LIMMA (AUC <0.40). These results also reflected in the Fig. 1b. On the other hand, for large-sample case in absence of outliers all the methods performed similarly. However, in presence of outliers in this case FCROS and KW performed better comparing with the others three methods (T-test, SAM and LIMMA).

To demonstrate the performance of all the methods in the real colon cancer dataset, firstly five methods were directly applied to identify the DE genes in this dataset. DE genes were detected using adjusted *p*-value <0.05. p-values were adjusted using Benjamini-Hochberg (1995) method. Fig. 3a represents the Venn diagram of DE genes detected by these five methods. From this Venn diagram, it was revealed that LIMMA and SAM performed better than the other three methods (T-test, KW and FCROS) by sharing more genes. The number of genes that were detected as DE by T-test, KW, SAM, LIMMA and FCROS, respectively is 367, 381, 352, 372 and 383. We also noticed that there are 274 genes common between these five methods. Fig. 3b shows the heat-map plot using hierarchical clustering analysis based on the common DE genes. There are 3, 59 and 14 genes that were identified by the T-test, FCROS and KW independently. Among these genes we explored the biological functions of 59 genes detected by FCROS using a web-based gene set analysis toolkit (Zhang et al. 2005). Fig. 4 represents the bar chart of the biological process, cellular component and molecular function categories. In this figure, biological Process, cellular component and molecular function category are represented by a red, blue and green bar, respectively. Using the KEGG database, it was found that these genes were involved in different pathways (Table 2 for top 10 pathways). In this table the hyper geometric test is used to calculate the p-values.

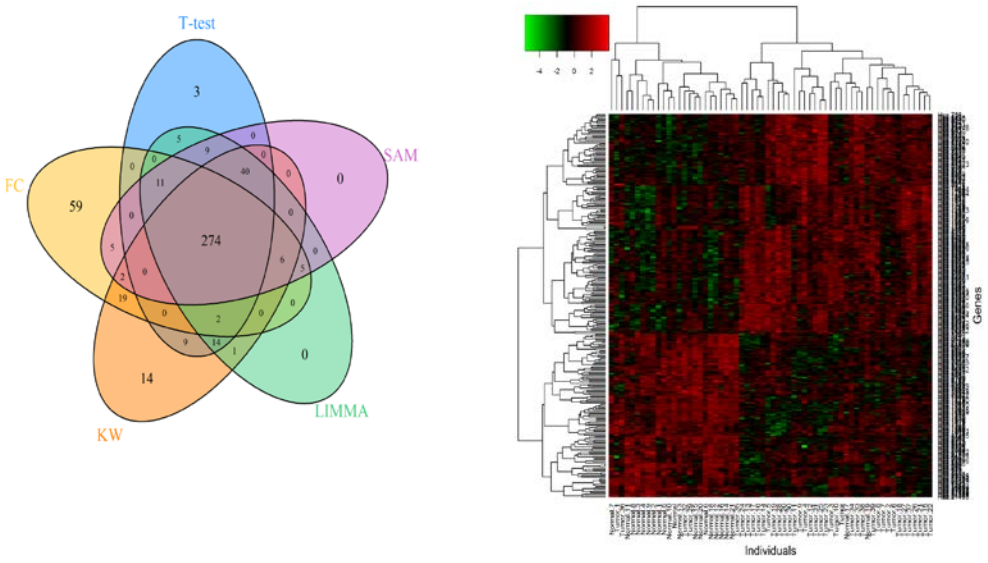

Fig. 3. Comparison of five methods for the selection of biomarker genes in colon cancer data set. (A) Venn diagram of DE genes detected by T-test, KW, SAM, LIMMA and FCROS, (B) Heat-map of 274 common DE genes detected by these five methods.

Table 2. Top 10 KEGG pathways for 59 DE genes detected by FCROS method for colon cancer data set.

KEGG ID	KEGG pathway names for *Homo sapiens* (human)	No. of genes	Adjusted -*p*-values
hsa00910	Nitrogen metabolism	10	7.02e-05
hsa00030	Pentose phosphate pathway	7	2.90e-03
hsa01230	Biosynthesis of amino acids	5	3.35e-05
hsa03018	RNA degradation	4	0.0002
hsa04931	Insulin resistance	3	0.0002
hsa01200	Carbon metabolism	3	0.0002
hsa05145	Toxoplasmosis	2	0.0001
hsa03040	Spliceosome	1	0.0001
hsa04630	Jak-STAT signaling pathway	1	0.0001
hsa05152	Tuberculosis	1	0.0001

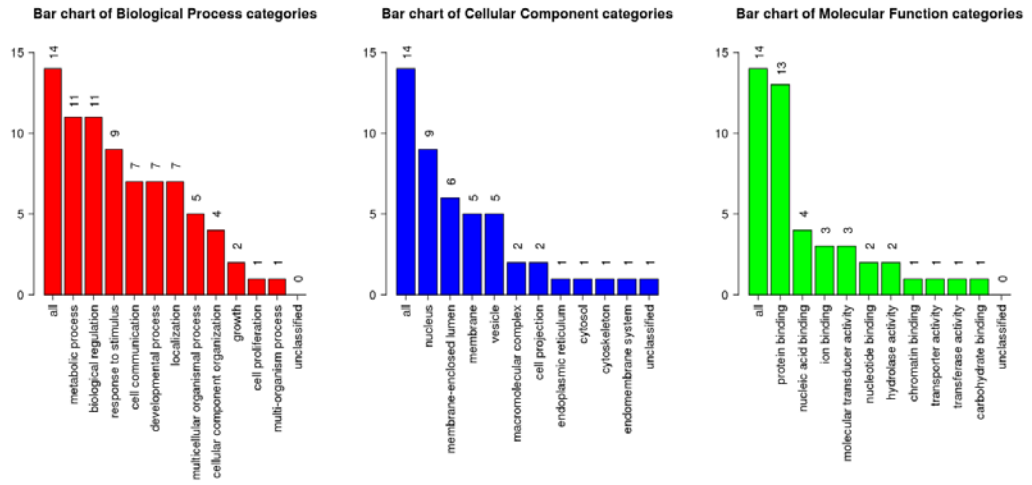

Fig. 4. Functional annotation of 59 DE genes identified by the FCROS method.

Conclusion

There are several methods existing in the literature to select biomarker gene and many comparative studies have been performed to select the appropriate method. But the choice of proper gene selection method is not easy when the gene expression dataset is contaminated by outliers and there is no specific guideline so far. In this paper, we evaluated the performance among the five popular gene selection methods using both simulated and real gene expression datasets in absence and presence of outliers. From the simulated data

analysis results, it was found that, in absence of outliers for both small-and large-sample cases, all the five methods performed well (except T-test for small-sample case). But in presence of outliers, for small-sample case FCROS method out performs other four methods. On the other hand, for large-sample case, in presence of outliers, KW and FCROS perform well. From a real colon cancer data analysis, it is elucidated that, FCROS method identified additional 59 genes that were not detected by the other methods. Using the KEGG pathway analysis, it is explored that this gene belongs to the different important pathways.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S and Mack D (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, USA 96 (12): 6745-6750.

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 57(1): 289-300.

Cui X and Churchill GA (2003). Statistical tests for differential expression in cDNA microarray experiments, Genome Biology 4(4): 1-10.

Dembele D and Kastner P (2014). Fold change rank ordering statistics: a new method for detecting differentially expressed genes, BMC Bioinformatics 15(14): 1-15.

Efron B, Tibshirani R, Goss V and Chu G (2001). Microarrays and their use in a comparative experiment, Journal of the American Statistical Association 96: 1151-1160.

Kaissi O, Nimpaye E, Singh TR,Vannier B, Ibrahimi A, Ghacham AA and Moussa A (2013). Genes selection comparative study in microarray data analysis, Bioinformation 9(20): 1019-1022.

Kerr MK, Martin M and Churchill GA (2000). Analysis of variance for gene expression microarray data, Journal of Computational Biology 7: 819-837.

Kruskal WH and Wallis WA (1952). Use of ranks in one-criterion variance analysis, Journal of the American Statistical Association 47: 583-621.

Mollah MMH, Jamal R, Mokhtar NM, Harun R and Mollah MNH (2015). A hybrid one-way ANOVA approach for the robust and efficient estimation of differential gene expression with multiple patterns, PLoS ONE 10(9): 1-26.

Shahjaman M, Kumar M, Mollah MMH, Ahmed MS, Begum AA, Islam SMS and Mollah MNH (2017). Robust significance analysis of microarrays by minimum β-divergence method, BioMed Research International 1-18.

Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). ROCR: visualizing classifier performance in R. Bioinformatics 21(20): 3940-3941.

Smyth GK, Thorne NP and Wettenhall J (2003). Limma: Linear Models for Microarray Data User's Guide. http://www.bioconductor.org.

Tibshirani R, Chu G, Narasimhan B and Li J (2013). SAM: Significance analysis of microarrays. http://cran.r-project.org/web/packages/samr/index.html.

Tusher V, Tibshirani R and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response, Proceedings of the National Academy of Sciences, USA 98: 5116-5121.

Zhang B Kirov S and Snoddy J (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts, Nucleic Acids Research 33: 741-748.