Asian J. Med. Biol. Res. 2020, 6 (4), 795-801; doi: 10.3329/ajmbr.v6i4.51248

Asian Journal of
Medical and Biological Research
ISSN 2411-4472 (Print) 2412-5571 (Online)
www.ebupress.com/journal/ajmbr

Article

A new statistical approach for the identification of outlier genes in cancer microarray data

Md. Bipul Hossen*

Department of Statistics, Faculty of Science, Begum Rokeya University, Rangpur, Rangpur-5404, Bangladesh

*Corresponding author: Md. Bipul Hossen, Department of Statistics, Faculty of Science, Begum Rokeya University, Rangpur, Rangpur-5404, Bangladesh. E-mail: mbipu@brur.ac.bd

Received: 07 December 2020/Accepted: 26 December 2020/ Published: 31 December 2020

Abstract: The aim of microarrays technology is to discover genes, which are differentially expressed as outliers between two or more groups of patients are an important task in the genomics community. The regular pattern of genes may often breakdown due to the presence of outliers and it is essential to detect those genes whose behavior looks abnormal in experimental and biological conditions. Several statistical techniques - t-statistic, cancer outlier profile analysis (COPA), outlier sums (OS), outlier robust t-statistic (ORT), maximum ordered subset t-statistics (MOST) and least sum of ordered subset square t-statistics (LSOSS) were developed to address the problem of detecting outlier genes in microarray data but these methods are affected by some problems especially if there is an unusual observation in such dataset then the standard assumptions of distribution parameter may be violated and these techniques might not be suitable to detect outliers genes as well. For these consequences, I have developed a new statistical technique that is “Propose t-statistic (PT)”. The performance of the newly proposed method PT statistic compare with the other existing methods applied to the monte carlo simulation data, package data, and real cancer datasets. The result shows that the outlier genes are identified by using the proposed method PT as well and will give the best and identical results than other methods. The performance of the proposed approach significantly improves than the traditional methods and it can extensively contribute to the medical as well as the genomic community.

Keywords: microarray data; monte carlo simulation; package data; outlier; t-test

1. Introduction

The last couple of decades are the revolution of omic technologies like genomics, transcriptomics, proteomics, and metabolomics, which played a significant contribution and drastic changes in all biomedical sciences and therapeutic management especially disease diagnosis, prognostication and development of novel therapies (Mazumdar and Raha, 2008). The fundamental key of analyzing microarray data is to identify genes in the context of experimental and phenotypical conditions. In particular, the objective of the study is to determine the genes that are outliers between the two classes of normal tissue versus tumor tissue (Griffiths *et al*., 1996; Hossen *et al*., 2018; Rashid *et al*., 2017).

Several statistical algorithms for this type of analysis have been proposed earlier. Initially, the fold changes simple method was used and it performance not so good because of its statistical variability was not taken into account stated by (Chen *et al*., 1997). After that, some other classical statistical methods have been proposed and these methods may not be reliable and may contain high noise data based on a single array microarray experiments. Since then the most widely used method student t-test which aims to check the statistical significance of every gene where the gene is not differentially expressed in different samples as under the null hypothesis (Phipson *et al*., 1963).

In microarray gene expression experiments, many methods were developed for aiming to detect cancer-related active genes. Recently, Tomlins (Tomlins *et al*., 2005) argued that classical and widely used analytical method two-sample t-statistic, which sometimes fails to detect cancer outlier genes that are over-expressed compared to normal genes. For this consequence, COPA (cancer outlier profile analysis) method were developed for finding the cancer genes tissue with dissimilar expression shapes within cancer tissue samples. Thereafter many statistical methods were proposed - Tibshirani and Hastie (2007) introduced the outlier sums (OS) method, Wu (2007) proposed the outlier robust t-statistic (ORT), Lian (2008) introduced the maximum ordered subset t-statistics (MOST) and a simple statistical test named least sum of ordered subset square t-statistic (LSOSS) is proposed by Wang and Rekaya (2010) for detecting cancer outlier differential gene expression.

However, most of the aforementioned statistical techniques for detection of outlier genes are not robust against outliers (Barnett and Lewis, 1994; Hawkins, 1980; Kannan and Monoj, 2015; Wang *et al.*, 2011), though it is a common problem in microarray data. Therefore, the main task is to detect outlier genes in the difficulties areas of biological conditions. To address this problem a new technique is proposed that is “Propose t-statistic (PT)” for detecting the outlier genes where it behavior shows highly different expression patterns in microarray data and compares this method with other existing methods. The helpfulness of the proposed method is then examined by Monte Carlo simulation data, package data and real gene expression cancer datasets.

2. Materials and Methods

2.1. Existing Outlier Detection Methods

2.1.1. t-statistic

Let x_{ij} be the values of expression for the genes i=1,2,3,…, m and samples j=1,2,3,…, n and consider that the samples are divided into two groups and group 1 denoted by normal or reference group, where group 2 is referred as disease group. The two sample t-statistic for gene i is defined as-

$$t_i = \frac{\bar{y}_i - \bar{x}_i}{s_i} \qquad (1)$$

Where $\bar{y}_i$ is the mean expression value in cancer tissues and is the mean expression value in normal tissues for gene i and s_i

$$s_i^2 = \frac{\sum_{1\le j\le n}(x_{ij} - \bar{x}_i)^2 - \sum_{1\le j\le m}(y_{ij} - \bar{y}_i)^2}{n+m-2} \qquad (2)$$

When most of the cancer tissues are activated the t-statistic is then powerful.

2.1.2. COPA (Cancer Outliers Profile Analysis statistic)

Tomlins *et al* (2005) proposed cancer profile outlier analysis (COPA) for detecting cancer outlier genes. They show that usual t-statistic is less powerful than COPA statistic in these cases.

It defines the COPA statistic as

$$copa_i = \frac{q_r\left(\{y_{ij}: 1 \le j \le m\}\right) - med_i}{mad_i} \qquad (3)$$

Where $\mathrm{q_r}$ (.) the rth, and $\mathrm{med_i}$ is the median expression value for each samples $\mathrm{med_i} = \mathrm{median}\left(\{\mathrm{x_{ij}}: 1 \le \mathrm{j} \le \mathrm{n}\}, \{\mathrm{y_{ij}}: 1 \le \mathrm{j} \le \mathrm{m}\}\right)$ and $\mathrm{mad_i}$ is the median absolute deviation that is $\mathrm{mad_i} = 1.4826 \times \mathrm{median}\left(\{\mathrm{x_{ij}} - \mathrm{med_i}\}: 1 \le \mathrm{j} \le \mathrm{m}\ \right)$.

2.1.3. Outliers Sums (OS) Statistic

Tibshirani and Hastie (2007) introduced a method called outlier sums for detecting the outlier’s genes. The limitation of COPA statistic is fixed rth sample percentile was considered by users. To overcome this problem OS statistic used and defined as

$$OS_i = \frac{\sum_{y_{ij}\in R_i}(y_{ij} - med_i)}{mad_i} \qquad (4)$$

Where,

$$R_i = \{y_{ij}: y_{ij} > q_{75}\left(\{x_{ij}: 1 \le j \le n\}, \{y_{ij}: 1 \le j \le m\}\right) + IQR\left(\{x_{ij}: 1 \le j \le n\}, \{y_{ij}: 1 \le j \le m\}\right)\}$$

and IQR(•) is the inter-quartile range $IQR\left(\{x_{ij}: 1 \le j \le n\}, \{y_{ij}: 1 \le j \le m\}\right) = q_{75}\left(\{x_{ij}: 1 \le j \le n\}, \{y_{ij}: 1 \le j \le m\}\right) - q_{25}\left(\{x_{ij}: 1 \le j \le n\}, \{y_{ij}: 1 \le j \le m\}\right)$

2.1.4. Outliers Robust t (ORT) statistic

ORT is usually similar to OS, and it uses the median value of normal sample instead of median of total data. Also estimates the absolute error using the median of several groups instead of the square error as in COPA for performing more robust and consistent estimation. Wu (2007) introduced this method as given-

$$ORT_i = \frac{\sum j \in R_i\,(\{y_{ij}\}_{1\le j\le n} - med_{ix}}{mad_i'} \qquad (5)$$

Where,$R_i = \{y_{ij}: y_{ij} > q_{75}(\{x_{ij}: 1 \le j \le n\}) + IQR(\{x_{ij}: 1 \le j \le n\})\}$ $mad_i' = 1.4826 \times median\ (\{(x_{ij} - med_{ix}): 1 \le j \le n\}, \{(y_{ij} - med_{iy}): 1 \le j \le m\})$

2.1.5. Maximum Ordered Subset t (MOST) statistic

To overcome the used of arbitrary outliers in OS and ORT statistic statistic Lian (2008) proposed MOST statistic because all possible values for outlier thresholds are considered. This statistic is calculated as

$$MOST_i = \begin{matrix} max \\ 1 \le k \le m \end{matrix}\left(\frac{\sum_{1\le j\le m}(y_{ij} - med_{ix})^2}{mad_i'} - \mu_k\right)/\delta_k \qquad (6)$$

Where k is unknown value, the data are normalized by μ_k and δ_k standard normal distribution.

2.1.6. The Least Sum of Ordered Subset Variance t-statistic (LSOSS)

When we check the distribution pattern of gene expression data it shows two peaks in cancer tissues. To address this problem of change of points or break down of peaks Wang and Rekaya (2010) proposed LSOSS statistic. Instead of median values, this method uses mean values in gene expression data and it defined as follows-

The expression values in cancer related tissues are ordered in magnitude and then divided into two groups for each gene i

$$S_{ik1} = \{y_{ij}: 1 \le j \le m\}, S_{ik2} = \{y_{ij}: k+1 \le j \le m\}$$

(a) Each gene i the mean and sum of squares are calculated for these two groups

$$\bar{y}_{S_{ik1}} = mean\{y_{ij}: 1 \le j \le m\}, \qquad \bar{y}_{S_{ik2}} = mean\{y_{ij}: k+1 \le j \le m\}$$

$$SS_{ik1} = \sum_{1\le j\le k}\left(y_{ij} - \bar{y}_{S_{ik1}}\right)^2, SS_{ik2} = \sum_{k+1\le j\le m}\left(y_{ij} - \bar{y}_{S_{ik2}}\right)^2$$

The remaining task is to solve the value of k which divides the two groups. By minimizing the pooled sum of squares the best value of k is determined for cancer tissues ranging from 1 to m – 1 are as-

$$arg \begin{matrix} min \\ 1 \le k \le m-1 \end{matrix}(SS_{ik1} + SS_{ik2})$$

Let S_{ix}^2 is the sum of squares amd S_i^2 be the pooled standard error for the estimated gene i for normal samples is given by

$\sum_{1\le j\le n}(x_{ij} - \bar{x}_i)^2$ and $S_i^2 = \frac{S_{ix}^2 + SS_{ik1} + SS_{ik2}}{n+m+2}$

(b) The LSOSS statistic for stating the outlier genes is computed as given below where k could be noted as the number of outlier samples for each gene i

$$LSSV_i = k\frac{\bar{y}_{S_{ik1}} - \bar{x}_i}{S_i} \qquad (7)$$

2.2. Proposed outlier detection method

In the earlier section of method discussion, most of the outlier gene detection techniques covers some non-robust statistic such as the mean and standard deviation and also observe that these statistic may not become effective to serve the purpose of detection genes. In this case, proposed new outlier technique developed to find the appropriate outlier genes and it described below-

2.2.1. Propose t-statistic (PT)

From the equation no (1) that t-statistic of two-condition for gene i is

$$t_i = \frac{\bar{y}_i - \bar{x}_i}{s_i}$$

Where $\bar{y}_i$ mean expression value in cancer tissues is, $\bar{x}_i$ is the mean expression value in normal tissue for gene i, since both $\bar{y}_i$ and $\bar{x}_i$ non-robust, therefore develop the propose t-statistic as

$$PT = \frac{med_{iy} - med_{ix}}{s_i} \qquad (8)$$

Where,

$med_{iy} = median\,(x_{iy}: 1 \leq j \leq m\,)$; $med_{ix} = median\,(x_{ix}: 1 \leq j \leq n\,)$ and

$$s_i^2 = \frac{\sum_{1\leq j\leq n}(x_{ij}-\bar{x}_i)^2 - \sum_{1\leq j\leq m}(y_{ij}-\bar{y}_i)^2}{n+m-2}$$

2.3. Data Sets

In this study colon cancer package datasets is used as a stable basis for checking performance and comparison of different method for real cancer gene expression data. The data sets analyzed in this study are in Table 1 and available at (Hossen *et al*., 2015).

3. Results and Discussion

Gene expression data can be detected outlier on both genes and samples and try to identify the outlier genes in microarray data based on the samples. The performance of the newly proposed method PT statistic compares with the other existing methods applied to the monte-carlo simulation data, package data and real gene expression cancer datasets.

3.1. Identification of outlier genes from monte carlo simulated data

Before applying to real data sets, I have to check the contribution of the newly proposed technique in simulation studies and compare it with the other existing methods. The simulation study was considered for different stages. The simulated data were generated from the standard normal distribution and generated 80 genes. Out of 80 genes, 50 genes were generated and consider the equal number of normal and cancer tissue samples with uniform condition for both groups. Remaining 30 genes were generated with two different conditions and assume that outliers do exist in these genes. The simulation process is taking for 7 times by changing the number of normal and cancer tissue sample sizes. In the first set of simulation n = 50 and m = 50 are generated as the number of samples from normal and cancer groups respectively. In other simulations, chose (n = 10, m =90), (n=35, m=65), (n=65, m=35), (n=75, m=25), (n=85, m=15) and (n=90, m=10) and the simulation results of the number of detected outlier genes are given in Table 2.

Table 2 shows that the result of the t-test performs well but in some situations, it fails to identify the original outliers. The methods of COPA, OS, ORT and MOST performance are not very satisfactory level rather than other methods. It also sees that the performance of LSSOS method gives the worst because most of the time it fails to detect not a single outlier properly. Besides, the proposed method PT performs best and gives identical results for all considered situations in simulation studies.

3.2. Identification of outlier genes from package data

The details of colon cancer package data sets consist of 2,000 genes where 40 tumor cancer tissues and 22 normal tissues exist (Merk, 1999). The ranges of tumor samples are 5.89 to 20903.18 whereas the ranges of normal samples are 5.82 to 14173.05. As I know that the genes with highly expressions pattern are identified as responsible for tumor tissue. For finding responsible tumor genes, I applied the classical outlier detection methods along with the newly proposed method. An indexed plot was considered to get an idea about tumor sample gene expression which is presented in Figure 1.

In Figure 1, the tumor sample genes are identified through cut-off points and the genes above the cut-off point 3 considered outlier genes. I get 27 common outlier genes that are detected all of the methods including the proposed method. Additionally, find the accuracy (the ratio of common genes and number of detected outlier genes) for all of the existing methods and proposed methods that are presented in Table 3.

Table 3 states the number of outliers detected by several methods along with accuracy level and show that the classical t-test can identify 39 genes as outliers. The COPA, OS, ORT, MOST and LSOSS statistic identify 47, 38, 52, 57 and 68 genes respectively. Whereas the newly proposed statistic PT can identify 32 genes with highest accuracy level than other methods.

3.3. Identification of Outlier Genes from Real Data

The newly proposed method and other outlier detection methods were applied to the six sets of real cancer datasets. The number of outliers is detected through the index plot at a certain cut-off points 3 for all methods. From the output genes (above cuff of points) try to calculate the common genes to check the accuracy of all methods. The common genes for the singh, golub v1, gordon, laiho, pomeroy-v1 and west data are 3, 6, 150, 150, 14 and 52 respectively. The number of the outlier and their corresponding accuracy are given in Table 4.

Table 4 shows that the LSOSS method cannot identify a single outlier in four datasets also the traditional t-test cannot identify a single outlier gene in two datasets whereas the COPA, OS, ORT, MOST, LSOSS and proposed statistic PT can identify outlier genes successfully. I also investigated the performance of proposed method compared with the existing methods through the accuracy measurement level and the results give that the accuracy of traditional t-test is 63.34%, and the accuracy of COPA, OS, ORT, MOST and LSOSS's are 26.90%, 26.67%, 30.21%, 26.24% and 22.38% respectively, whereas the accuracy level of PT is 70.17%. The LSOSS method gives the worst results followed by the OS, COPA, MOST, ORT and t-test methods. The Proposed method PT which can identify the outlier genes successfully and achieve the highest accuracy level in all datasets except Laiho data. So I may conclude that the proposed method gives the best results among all other methods.

Table 1. Data Description.

Dataset	Chip	Tissue	Sample No.	Dist. Classes	Gene
Golub-V1	Affy	Bone marrow	72	n=47, m=25	1877
Gordon	Affy	Lung	181	n=31, m=150	1626
Pomeroy-V1	Affy	Brain	34	n=25, m=9	857
Laiho	Affy	Colon	37	n=8, m=29	2202
Singh	Affy	Prostate	102	n=50, m=52	339
West	Affy	Breast	49	n=25, m=24	1198

Table 2. Outliers in different methods in Simulation Study.

Methods	n=35 m=65 gene=40	n=65 m=35 gene=40	n=75 m=25 gene=40	n=50 m=50 gene=40	n=10 m=90 gene=40	n=90 m=10 gene=40	n=85 m=15 gene=40
T	18	19	20	18	17	20	20
COPA	14	11	14	16	16	8	3
OS	13	12	11	14	0	5	7
ORT	15	12	11	17	17	3	7
MOST	16	13	10	15	16	9	4
LSOSS	25	0	0	0	26	0	0
PT	**20**	**19**	**20**	**20**	**20**	**20**	**19**

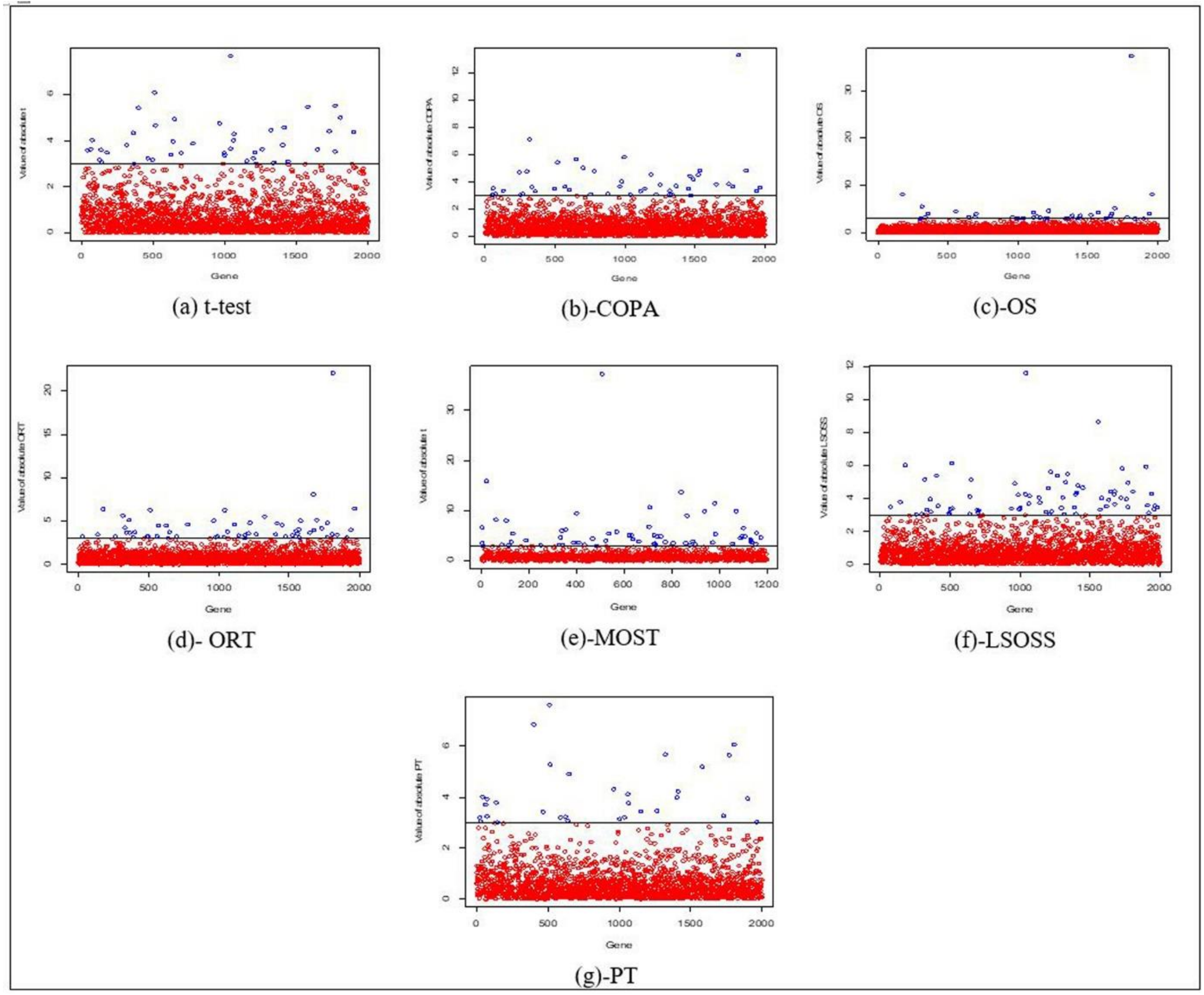
Table 3. Number of Detected Outliers for Package Data.

Method	Number of Outliers (n = 22, m = 40, g=2000)	Accuracy
t-test	39	69.23%
COPA	47	57.44%
OS	38	71.05%
ORT	52	51.92%
MOST	57	47.36%
LSOSS	68	39.70%
PT	**32**	**84.37%**

Table 4. Summery for the accuracy value of real cancer datasets.

Methods	Singh		Golub V1		Gordon		Laiho		Pomeroy-V1		West		Av. Ac
	NO	Ac	NO	Ac	NO	Ac	NO	Ac	NO	Ac	NO	Ac	
t-test	6	50%	18	33.33%	0	0	0	0	16	87.5%	63	82.54%	63.34%
COPA	14	21.43%	456	1.316%	273	54.95%	227	66.07%	133	10.52%	303	17.16%	26.90%
OS	14	21.43%	530	1.113%	310	48.38%	216	69.45%	325	4.30%	338	15.38%	26.67%
ORT	27	11.11%	497	1.22%	254	59.06%	**179**	**83.79%**	324	4.32%	240	21.67%	30.21%
MOST	14	21.43%	264	2.27%	315	52.38%	264	56.81%	214	6.54%	289	17.99%	26.24%
LSOSS	14	21.43%	0	0	0	0	0	0	60	23.33%	0	0	22.38%
PT	**5**	**60%**	**16**	**37.5%**	**165**	**90.91%**	302	49.66%	**15**	**93.33%**	**58**	**89.67%**	**70.17%**

[N.B.; NO= Number of Outlier, Ac= Accuracy and it calculate by the ratio of common outlier genes and total detected outlier genes by the methods, Av. Ac= Average accuracy and it calculate the mean value of all datasets accuracy.]



(a) t-test

(b)-COPA

(c)-OS

(d)- ORT

(e)-MOST

(f)-LSOSS

(g)-PT

Figure 1. Index plot of genes with cut-off points for existing outlier detection methods and proposed method of colon cancer package data.

4. Conclusions

In microarray gene expression data, discovering the highly expressed genes of the cancer tissue sample is an important growing field of medical research because it can optimize gene expression for cell growth for which it is easy to understand which tissue samples are involved or responsible for cancerous tissues. The performances of proposed statistic PT are better than other existing methods in simulation results suggest that. For package data, proposed methods PT can identify 32 genes as outlier genes among 27 common outlier genes whereas the existing method could not perform as well. For real cancer gene expression datasets, the average accuracy of the existing methods are respectively for t-statistics, COPA, OS, ORT, MOST and LSOSS are 63.34%, 26.90%, 26.67%, 30.21%, 26.24% and 22.38%. Whereas the proposed method PT will give the highest accuracy of 70.17%. Therefore I may conclude that my proposed method which is used to identify the outliers genes in the simulated data, package data and gene expression microarray real dataset gives the best and identical results than other methods. This method will afford useful information, which can help to develop the performance and could assist to detect outlier genes.

Conflict of interest

None to declare.

References

Barnett V and T Lewis, 1994. Outliers in Statistical Data, John Wiley.

Chen Y, ER Dougherty and ML Bittner, 1997. Ratio based decisions and the quantitative analysis of cDNA microarray images. J. Biomedical Optics, 2: 364-367.

Griffiths JF, JH Miller, DT Suzuki, RC Lewontin and WM Gelbart, 1996. An Introduction to Genetic Analysis, W. H. Freeman and Company. New York: 6th edition.

Hawkins D, 1980. Identification of Outliers. Chapman and Hall.

Hossen MB, Siraj-Ud-Doulah and A Hoque, 2015. Methods for evaluating agglomerative hierarchical clustering for gene expression data: a comparative study. Computational Biology and Bioinformatics, 3: 88-94.

Hossen MB and Siraj-Ud-Doulah, 2017. Identification of robust clustering methods in gene expression data analysis. Current Bioinformatics, 12: 558-562.

Kannan KS and K Manoj, 2015. Outlier Detection in Multivariate Data. Applied Mathematical Sciences, 9: 2317-2324.

Lian H, 2008. MOST: detecting cancer differential gene expression. Biostatistics, 9:411–431.

Mazumdar D and S Raha, 2008. Evolution to revolution; a review on bioinformatics. Advanced modelling and Optimization, 10: 51-62.

Merk S. colonCA: exprSet for Alon *et al*. 1999. Colon cancer data. R package version 1.22.0. 2018.

Phipson B, S Lee, IJ Majewski, WS Alexander and GK Smyth, 2016. Robust hyper parameter estimation protects against hyper variable genes and improves power to detect differential expression. Ann. Appl. Stat., 10: 946–963.

Rashid or Harun, A Mowla, S Rahman, Siraj-Ud-Doulah and MB Hossen, 2017. Statistical tests for identification of differentially expressed genes in microarray data. Biomedical Statistics and Informatics, 2: 166-171.

Tibshirani R and T Hastie, 2007. Outlier sums for differential gene expression analysis. Biostatistics, 8: 2–8.

Tomlins SA, DR Rhodes, S Perner, SM Dhanasekaran, R Mehra, Xiao-Wei Sun, S Varambally, X Cao, J Tchinda, R Kuefer, C Lee, JE Montie, RB Shah, KJ Pienta, MA Rubin and AM Chinnaiyan, 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science, 310: 644–648.

Wang Y and R Rekaya, 2010. LSOSS: Detection of Cancer Outlier Differential Gene Expression. Biomarker Insights, 5: 69–78.

Wang Y, C Wu, Z Ji, B Wang and Y Liang, 2011. Non-parametric change-point method for differential gene expression detection. PLoS ONE, 6: e20060.

Wu B, 2007. Cancer outlier differential gene expression detection. Biostatistics, 8: 566–575.