Robust and efficient identification of biomarkers from RNA-Seq data using median control chart [version 1; peer review: 1 approved with reservations, 1 not approved]

Md Shahjaman 1, Habiba Akter 1, Md. Mamunur Rashid 1, Md. Ibnul Asifuzzaman 1, Md. Bipul Hossen 1, Md. Rezanur Rahman 2

1Department of Statistics, Begum Rokeya University, Rangpur, 5400, Bangladesh
2Department of Biochemistry and Biotechnology, Khwaja Yunus Ali University, Sirajgonj, 6200, Bangladesh

Abstract

Background: One of the main goals of RNA-seq data analysis is identification of biomarkers that are differentially expressed (DE) across two or more experimental conditions. RNA-seq uses next generation sequencing technology and it has many advantages over microarrays. Numerous statistical methods have already been developed for identification the biomarkers from RNA-seq data. Most of these methods were based on either Poisson distribution or negative binomial distribution. However, efficient biomarker identification from discrete RNA-seq data is hampered by existing methods when the datasets contain outliers or extreme observations. Specially, the performance of these methods becomes more severe when the data come from a small number of samples in the presence of outliers. Therefore, in this study, an attempt is made to propose an outlier detection and modification approach for RNA-seq data to overcome the aforesaid problems of traditional methods. We make our proposed method facilitate in RNA-seq data by transforming the read count data into continuous data.

Methods: We use median control chart to detect and modify the outlying observation in a log-transformed RNA-seq dataset. To investigate the performance of the proposed method in absence and presence of outliers, we employ the five popular biomarker selection methods (edgeR, edgeR_robust, DEseq, DEseq2 and limma) both in simulated and real datasets.

Results: The simulation results strongly suggest that the performance of the proposed method improved in the presence of outliers. The proposed method also detected an additional 18 outlying DE genes from a real mouse RNA-seq dataset that were not detected by traditional methods. Using the KEGG pathway and gene ontology analysis results we reveal that these genes may be biomarkers, which require validation in a wet lab.

Conclusions: Our proposal is to apply the proposed method for biomarker identification from other RNA-seq data.
Keywords
RNA-Seq data, Logarithmic transformation, Biomarker, Outliers and Robustness

Corresponding authors: Md Shahjaman (shahjaman_br@yahoo.com), Habiba Akter (habiba23bru@gmail.com)

Author roles: Shahjaman M: Conceptualization, Supervision, Writing – Review & Editing; Akter H: Methodology; Rashid MM: Resources, Software; Asifuzzaman Mi: Conceptualization, Writing – Review & Editing; Hossen MB: Writing – Review & Editing; Rahman MR: Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2019 Shahjaman M et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Shahjaman M, Akter H, Rashid MM et al. Robust and efficient identification of biomarkers from RNA-Seq data using median control chart [version 1; peer review: 1 approved with reservations, 1 not approved] F1000Research 2019, 8:7 (https://doi.org/10.12688/f1000research.17351.1)

First published: 03 Jan 2019, 8:7 (https://doi.org/10.12688/f1000research.17351.1)
Introduction

One of the major objectives of researchers is to identify biomarkers from RNA-Seq data that are differentially expressed (DE) between two or more experimental conditions. Microarrays have been extensively used during the past few decades to perform this task. But after reducing the cost of sequencing, biomarker identification using RNA-seq data has emerged as an alternative choice to microarrays. RNA-seq uses next generation sequencing technology to produce a vast amount of data. *Curse of dimensionality* is a common problem for analyzing RNA-Seq data, which means a “large p, small n” problem. Hence, dimension reduction of the data matrix is a primary objective for further downstream analysis. Identification of biomarkers is one of the dimensionality reduction approaches.

RNA-seq count data inherently follow a Poisson or negative binomial (NB) distribution rather than normal distribution like microarray data. Numerous statistical methods have been developed to identify biomarkers from RNA-seq count data. The earliest method is DESeq, which is based on Poisson distribution. This method suffers from overdispersion and therefore Poisson distribution-based methods are not suitable for RNA-seq data. To overcome this problem, NB distribution-based methods have been proposed. Some NB based methods are: baySeq, DESeq, DESeq2, EBSeq, edgeR, edgeR (robust) and NBSeq. However, most of the methods cannot estimate properly the gene-wise dispersion parameters and they also suffer from small sample sizes. The RNA-Seq count data can be equal to zero, so we shift them by one before transforming them:

\[ x_{gik} = \log(y_{gik} + 1) \]

Log-transformed values have less extreme values (or outliers) than the untransformed data.

Despite the popularity of these statistical methods for identification of biomarker genes, they are sensitive to outliers and often produce lower accuracies in the presence of outliers. Outliers may arise in RNA-seq count data because there are several data generating stages from biological harvesting of RNA samples to counting of sequence read map data. To mitigate this issue many algorithms use transformation methods. There are several transformation methods for RNA-seq data: logarithmic transformation, variance-stabilizing transformation (vst), TMM transformation, regularized logarithm and variance modeling at the observation level (voom). These methods only reduce the low level outliers into reasonable spaces during parameter estimations; however they fail to reduce the influence of high level outliers with small sample sizes in the data matrix.

Consequently, most biomarker selection methods that use the aforesaid transformation methods, are sensitive to outliers or extreme values with small-sample sizes. Therefore, in this study, an attempt is made to propose an outlier detection and modification approach for RNA-seq data to improve the performance of the popular biomarker selection methods in the presence of outliers. To make our proposed method facilitate in RNA-seq data we transform the read count data into continuous data using regularize logarithmic transformation.

The article is organized as follows: *Methods* briefly describes the logarithmic transformation and formulation the proposed outlier detection and modification approach. In *Results and Conclusions* a broad simulation study and a real data study have been carried out.

Methods

Let \( y_{gik} \) be the number of reads simulated from \( g \)th gene of \( k\)th replicate in the \( i\)th condition (\( g = 1, 2, \ldots , G; i = 1, 2, \ldots, n \)); where \( n \) is the number of replications in condition \( i \). \( G \) is the total number of gene. Then the negative binomial distribution is as follows:

\[ y_{gik} \sim \text{Negative Binomial} \left( \mu_{g_i}, \text{size} = r_{g_i} \right) \]

In this negative binomial parameterization, \( E(y_{gik}) = \mu_{g_i} \) and \( \text{Var} \ (y_{gik}) = \mu_{g_i} + \frac{\mu_{g_i}^2}{r_{g_i}} \) where, \( \mu_{g_i} \) is the mean of \( g \)th gene in \( i\)th condition and \( \frac{1}{r_{g_i}} \) is the dispersion parameter. Now we want to test the following null hypothesis:

\[ H_0 : \mu_{g1} - \mu_{g2} = 0 \quad \text{Vs} \quad H_1 : \mu_{g1} \neq \mu_{g2} \]

A gene will be declared as DE if \( H_0 \) is rejected, otherwise it is equally expressed (EE).

Log-transformation

Log-transformation is very useful in RNA-Seq data. The log-transformed data usually follow the normal distribution, which depends on the degree of skewness before transformation. As the RNA-Seq count data can be equal to zero, so we shift them by one before transforming them:

\[ x_{gik} = \log(y_{gik} + 1) \]

Let \( x_{gik} \) be the number of reads simulated from \( g \)th gene of \( k\)th replicate in the \( i\)th condition (\( g = 1, 2, \ldots , G; i = 1, 2, \ldots, n \); where \( n \) is the number of replications in condition \( i \). \( G \) is the total number of gene. Then the negative binomial distribution is as follows:

\[ y_{gik} \sim \text{Negative Binomial} \left( \mu_{g_i}, \text{size} = r_{g_i} \right) \]

In this negative binomial parameterization, \( E(y_{gik}) = \mu_{g_i} \) and \( \text{Var} \ (y_{gik}) = \mu_{g_i} + \frac{\mu_{g_i}^2}{r_{g_i}} \) where, \( \mu_{g_i} \) is the mean of \( g \)th gene in \( i\)th condition and \( \frac{1}{r_{g_i}} \) is the dispersion parameter. Now we want to test the following null hypothesis:

\[ H_0 : \mu_{g1} - \mu_{g2} = 0 \quad \text{Vs} \quad H_1 : \mu_{g1} \neq \mu_{g2} \]

A gene will be declared as DE if \( H_0 \) is rejected, otherwise it is equally expressed (EE).

Biomarker identification using the proposed procedure

Since the median and median absolute deviation is the robust measure of location and scale parameter, respectively, therefore, we used the median control chart as a measure of outlier detection. The proposed procedure is as follows:

1. We declare a gene as an outlying gene if it doesn’t fall into the interval \([LCL, UCL]\). Where LCL and UCL are the lower and upper control limit for median and they are defined by \( \text{LCL} = \text{MED}_{g} - 3 \times \text{MAD}_{g} \) and \( \text{UCL} = \text{MED}_{g} + 3 \times \text{MAD}_{g} \). Here, \( \text{MED}_{g} \) is median

\[ (x_{gik}; \ g = 1, 2, \ldots , G; \ k = 1, 2, \ldots, n; \ t = 1,2) \]

is the median of \( g \)th gene in \( t\)th condition, \( \text{MAD}_{g} = \text{median}_{i=1,2,\ldots,n}(|x_{gik} - \text{MED}_{g}|) \) is the median absolute deviation.

2. Check the existence of outliers for each gene from each of the conditions \((i=1,2)\), separately using step 1. If outliers exist, replace them by their respective group medians (\( \text{MED}_{g} \)).
3. Apply the anti-log transformation to obtain modified RNA-seq (MRS) count datasets.

4. Apply traditional statistical methods in MRS data to identify biomarker genes using the p-values adjusted by Benjamini-Hochberg method.

5. Obtain the functional annotations and KEGG pathways for detected biomarker genes.

Performance evaluation

In order to evaluate the performance of different biomarkers selection methods we considered the area under the receiver operating characteristic curve (ROC) curve. The ROC is created by plotting the true positive rate (TPR) against the false positive rate (FPR) for different cut-off points of a parameter. For a particular threshold each point on the ROC curve produces a TPR/FPR pair. The area under the ROC curve (AUC) is a performance measure which helps us to select an optimum method that can distinguish between two gene groups such as DE or EE well.

Datasets

To investigate the performance of the proposed method in comparison with five popular methods as mentioned above, for both small-and-large-sample cases with 2 groups/conditions, we considered 100 datasets for both cases with sample sizes of \(n_1=n_2=3\) and 15, respectively. Each dataset for each case represented gene expression profiles for 1000 genes, each with \(n=(n_1+n_2)\) samples, where the read counts of each gene was generated using the negative-binomial distribution and this type of simulation study was also used in 11. The number of DE genes were set to 40 for each of the 100 datasets. We divided these 40 DE genes into two groups: 20, up-regulated DE genes and 20, down-regulated DE genes. To show the effect of outliers (extreme high counts) on the methods, we randomly selected 10% and 30% genes and for each of these genes, we selected a single sample randomly and multiplied the observed count of this sample with randomly selected factor between 5 and 10. This process was applied for each of the 100 datasets. We computed average values of different performance measures such as true positive rate (TPR), false positive rate (FPR) and AUC based on 40 estimated DE genes by five methods (edgeR, edgeR_robust, DESeq, DESeq2, and limma (voom)) for each of 100 original datasets and proposed MRS datasets.

We also considered a real RNA-seq mouse dataset\(^7\) to demonstrate the performance of the methods. This dataset consists of 36535 genes with 21 samples. This dataset was downloaded from ReCount website http://bowtie-bio.sourceforge.net/recount. It can also be downloaded from the GEO series accession number GSE26024. Among 21 samples, RNA-seq count expression collected from 10 C57BL/6J (B6) and 11 DBA/2J (D2) inbred mouse strains.

Software

To demonstrate the performance of the proposed method, a comparison with five popular methods (edgeR, edgeR_robust, DESeq, DESeq2 and limma (voom)) was performed. We used both simulated and real RNA-seq count datasets. We used three R packages of other methods: edgeR, DESeq and limma. The performance measure area under the receiver operating characteristics curve (AUC) was computed for each of the methods using R package ROCR. All R packages are available in the comprehensive R archive network (cran) or Bioconductor.

Results

Performance evaluation based on simulated dataset

Table 1 summarizes the average AUCs estimated by eight methods based on 100 simulated datasets using 4% DE genes in absence and presence of single outlier in each of 10% and 30%.

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Outliers</th>
<th>edgeR</th>
<th>edgeR (robust)</th>
<th>DESeq</th>
<th>DESeq2</th>
<th>Limma (voom)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Outliers</td>
<td>0.844</td>
<td>(0.844)</td>
<td>0.851</td>
<td>(0.85)</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>10% Outliers</td>
<td>0.829</td>
<td>(0.842)</td>
<td>0.826</td>
<td>(0.825)</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>30% Outliers</td>
<td>0.697</td>
<td>(0.796)</td>
<td>0.709</td>
<td>(0.786)</td>
<td>0.713</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Outliers</th>
<th>edgeR</th>
<th>edgeR (robust)</th>
<th>DESeq</th>
<th>DESeq2</th>
<th>Limma (voom)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Outliers</td>
<td>0.966</td>
<td>(0.965)</td>
<td>0.943</td>
<td>(0.942)</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>10% Outliers</td>
<td>0.940</td>
<td>(0.959)</td>
<td>0.934</td>
<td>(0.946)</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td>30% Outliers</td>
<td>0.894</td>
<td>(0.937)</td>
<td>0.907</td>
<td>(0.949)</td>
<td>0.852</td>
</tr>
</tbody>
</table>
genes for both small-sample cases ($n_1=n_2=3$) and large-sample cases ($n_1=n_2=15$), respectively. In Table 1, the results without and within the brackets (.) indicate the estimated AUCs by the five methods using the original RNA-seq datasets and proposed MRS datasets. From Table 1 we observed that in absence of outliers, five methods (edgeR, edgeR_robust, DESeq, DESeq2 and limma (voom)) produced almost similar results using the original RNA-seq datasets and proposed MRS datasets, for both small-and-large-sample cases. However, in the presence of outliers, the performance of these methods has significantly increased using the proposed MRS datasets for both cases. For example, in the presence of 10% outliers, edgeR and DESeq produce AUCs 0.829 and 0.818, respectively for small-sample case. Whereas for the same condition these two methods produce AUCs 0.842 and 0.838, respectively using our proposed MRS datasets. Figure 1a and b and Figure 1c and d show the box-plot of AUCs based on 100 simulated datasets by each of the methods in absence and presence of outliers for small-and large-sample cases, respectively. The left and right-side panels in this figure indicate the boxplot of estimated AUCs using original RNA-seq datasets and proposed MRS datasets. Similar results were found from these boxplots, as in Table 1.

**Performance evaluation based on real dataset**

After filtering we retain 11,474 genes. To investigate the performance of the proposed method, we employ three methods (edgeR, DESeq and limma) for detection of the biomarkers between the two mouse strains. Figure 2a and b represents the Venn diagram of estimated DE genes by edgeR, DESeq and Limma using the original and MRS dataset, respectively. From Figure 2a, we revealed that edgeR and Limma performed better than DESeq by sharing more genes (414). We also noticed that there are 1925 overlapping DE genes between these methods. To investigate the performance of the proposed outlier detection and modification approach in this dataset, we first detect and modify the outliers (if any) to get the MRS dataset. We detected 200 outliers in this dataset using the proposed outlier detection rule. The Venn diagram in Figure 2b represents the results of these three methods using the proposed MRS dataset. From this figure we can clearly observe that there are 1956 overlapping

![Figure 1](image)

**Figure 1.** Performance evaluation using boxplot of AUCs estimated by different methods using the simulated datasets. (a-b) for small-sample case ($n_1=n_2=3$) and (c-d) for large-sample case ($n_1=n_2=15$).
Figure 2. Comparison of the DEGs detected by three methods and outlying gene expression profiles for Mouse dataset. Venn diagram of DEGs detected by (a) the edgeR, DESeq and Limma in the original mouse dataset or by (b) the edgeR, DESeq and Limma in the modified mouse dataset using the proposed method. (c) Heatmap of 16 outlying DEGs detected by the proposed method.

Among these genes there are 18 genes that are declared as outliers by the proposed method and those were not detected as DE genes using the original mouse dataset.

Furthermore, we performed the gene overexpression analyses through Database for Annotation, Visualization and Integrated Discovery (DAVID)\(^8\) to explore the biological process (BF) categories and pathway annotations of the 18 identified outlying DE genes. Out of 18 genes DAVID identified 16 genes. A heatmap is created for these 16 outlying DE genes in Figure 2c. The heatmap correctly clusters the samples between C57BL/6J (B6) and 11 DBA/2J (D2) using these genes. Among the 16 genes, 8 upregulated (Fam46b, Alx3, Dusp2, Pdyn, Agbl2,
Pcdh12, Ubl5, Gpx8) and 8 downregulated (Stard5, Ptprc, Slc7a5, Slc24a1, Ehd2, Adgrg3, Tefm, Tsnaxip1) DE genes are identified. GO analysis results showed that upregulated DE genes are significantly enriched in protein side chain deglutamylation, protein deglutamylation and embryo development at BP level. Downregulated DEGs were enriched in negative regulation of CREB transcription factor activity, negative regulation of NIK/NF-kappaB signaling and sterol import at BP level (see extended data). KEGG analysis showed that the upregulated DEGs were mostly enriched in cocaine addiction, glutathione metabolism and thyroid hormone synthesis. The downregulated DEGs were enriched in phototransduction, primary immunodeficiency and central carbon metabolism in cancer (Table 2). We also constructed the protein-protein interaction (PPI) network around the proteins encoded by these 16 outlying genes using STRING database. We considered confidence score 400 for selection these networks. Figure 3 and Figure 4 represent the PPI networks using the up-regulated and down-regulated DEGs, respectively. In addition, we explored miRNAs-target gene interactions from miRTarBase to identify miRNAs. The miRNAs-target gene interactions network is shown in Figure S1 in extended data.

Conclusions
Biomarker identification under two or more conditions is an important task for elucidating the molecular basis of phenotypic variation. Next generation sequencing (RNA-seq) has become very popular and a competitive alternative to microarrays because of reducing the cost of sequencing and limitation of microarrays. A number of methods have been developed for detecting biomarkers from RNA-seq data. However, most of the methods are sensitive to outliers and produce misleading results in the presence of outliers. In this study, we have proposed an outlier detection and modification approach using the median control chart. From the simulation study in the presence of outliers we have observed that the performance of five biomarker selection methods are improved significantly when the datasets are modified by the proposed method, both for small-and large-sample cases. The proposed method also detected an additional 16 outlying genes from a real mouse dataset. From GO and KEGG pathway

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Pathway Name</th>
<th>Name of Gene</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmu05030</td>
<td>Cocaine addiction</td>
<td>Pdyn</td>
<td>1.79e-02</td>
</tr>
<tr>
<td>mmu00480</td>
<td>Glutathione metabolism</td>
<td>Gpx8</td>
<td>2.29e-02</td>
</tr>
<tr>
<td>mmu05031</td>
<td>Amphetamine addiction</td>
<td>Pdyn</td>
<td>2.53e-02</td>
</tr>
<tr>
<td>mmu04918</td>
<td>Thyroid hormone synthesis</td>
<td>Gpx8</td>
<td>2.79e-02</td>
</tr>
<tr>
<td>mmu00590</td>
<td>Arachidonic acid metabolism</td>
<td>Gpx8</td>
<td>3.31e-02</td>
</tr>
<tr>
<td>mmu05034</td>
<td>Alcoholism</td>
<td>Pdyn</td>
<td>7.29e-02</td>
</tr>
<tr>
<td>mmu04010</td>
<td>MAPK signaling pathway</td>
<td>Dusp2</td>
<td>9.17e-02</td>
</tr>
<tr>
<td>mmu04744</td>
<td>Phototransduction</td>
<td>Slc24a1</td>
<td>1.35e-02</td>
</tr>
<tr>
<td>mmu05340</td>
<td>Primary immunodeficiency</td>
<td>Ptprc</td>
<td>1.79e-02</td>
</tr>
<tr>
<td>mmu05230</td>
<td>Central carbon metabolism in cancer</td>
<td>Slc7a5</td>
<td>3.27e-02</td>
</tr>
<tr>
<td>mmu04666</td>
<td>Fc gamma R-mediated phagocytosis</td>
<td>Ptprc</td>
<td>4.38e-02</td>
</tr>
<tr>
<td>mmu04660</td>
<td>T cell receptor signaling pathway</td>
<td>Ptprc</td>
<td>5.16e-02</td>
</tr>
<tr>
<td>mmu04150</td>
<td>mTOR signaling pathway</td>
<td>Slc7a5</td>
<td>7.59e-02</td>
</tr>
<tr>
<td>mmu04514</td>
<td>Cell adhesion molecules (CAMs)</td>
<td>Ptprc</td>
<td>8.2e-02</td>
</tr>
<tr>
<td>mmu04144</td>
<td>Endocytosis</td>
<td>Ehd2</td>
<td>1.36e-01</td>
</tr>
</tbody>
</table>
enrichment analysis, we have shown that these genes belong to some important pathways.

Data availability
Underlying data
Simulated datasets available from: https://doi.org/10.5281/zenodo.2212881

Real dataset: The mouse dataset used in this study is publicly available at the NCBI GEO website: GSE26024.

Extended data
Zenodo: Figure S1: miRNAs-target gene interactions using the outlying genes identified by the proposed method, http://doi.org/10.5281/zenodo.2279921

Software availability
The R code for the proposed method is available in https://github.com/snotjanu/OutMod-RnaSeq
Archived code: http://doi.org/10.5281/zenodo.227940
License: MIT

References


Grant information
The author(s) declared that no grants were involved in supporting this work.

Acknowledgements
We would like to thank the reviewers for their valuable comments on the paper, as these comments led us to an improvement of the work.
Open Peer Review

Current Peer Review Status:  ?  

Version 1

Reviewer Report 06 June 2019

https://doi.org/10.5256/f1000research.18975.r48899

© 2019 Li J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jun Li
Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA

The author proposed a method for outlier detection and differential expression (DE) identification for RNA-seq data. While DE is surely an important problem in RNA-seq data analysis, the proposed method is based on a wrong mathematical model and thus makes no sense.

The author assumes the read count $y_{gik}$ follows a negative binomial model with mean $\mu_{gi}$ depending only on the gene ($g$) and the condition ($i$). Where is the sequencing depth? Normalizing/incorporating sequencing depth has been a central question in DE analysis and significant efforts have been made by many important papers in this field, but the author completely ignored this term. Similarly, without considering the sequencing depth, the null hypothesis the author wrote is also wrong. With the wrong model and wrong hypothesis for testing, the proposed method does not make any sense.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
No

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes
Are the conclusions drawn adequately supported by the results?
No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** statistics, biostatistics, bioinformatics, sequencing data analysis

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 29 April 2019
https://doi.org/10.5256/f1000research.18975.r47416

© 2019 Li L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lei Li
Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine (BCM), Houston, TX, USA

In the manuscript, Shahjaman et al. presents a new median control chart method for controlling outliers in RNA-Seq data. They applied this method on simulated data and also a real mouse dataset. In general, a useful bioinformatics method for outlier detection will be valuable for biomarker detection. However, in the present state of the manuscript the robust and efficiency of this method is not clear. There are a few concerns that should be considered to improve the manuscript.

1. Outlier methods has actually been implemented in a few tools such as DESeq2 (https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#appr)
   The authors need to compare their method with other available outlier detection methods? Also, what is the tool parameters for the authors use in their evaluation? The outlier performance may largely depend on the parameters used.
2. The figure legend needs to be well described. What is the meaning of Figure 1A and Figure 1C, is that one for pre-control and one for after-medial control? Also, Figure 1c and Figure 1d?
3. Why only use three, not five methods for real data performance evaluation?
4. The released github code cannot reproduce the authors' result.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes
Are all the source data underlying the results available to ensure full reproducibility? 
Partly

Are the conclusions drawn adequately supported by the results? 
Partly

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* RNA seq

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com