Exploratory Analysis of Neuroblastoma Data Genes Expressions Based on Bioconductor Package Tools

Sergii Babichev[1,2][0000-0001-6797-1467], Bohdan Durnyak[2][0000-0003-1526-9005], Vsevolod Senkivskyy[2][0000-0002-4510-540X], Oleksandr Sorochynskyi[2][0000-0003-0964-2598], Mykhailo Kliap[3][0000-0003-1933-6148] and Orest Khamula[2][0000-0003-0926-9156]

[1]Jan Evangelista Purkyne University in Usti nad Labem, Usti nad Labem, Czech Republic
sergii.babichev@ujep.cz
[2]Ukrainian Academy of Printing, Lviv, Ukraine
durnyak@uad.lviv.ua, senk.vm@gmail.com, somsoroka@gmail.com
[3]Uzhhorod National University, Uzhhorod, Ukraine
mihaylo.klyap@uzhnu.edu.ua

Abstract. The technique of gene expression profiles exploratory analysis on the basis of the use of Bioconductor package tools is presented in the paper. Applying this method allows forming the matrix of genes expression for further gene regulatory networks reconstruction and simulation of the obtained model. The gene expression profiles of human neuroblastoma cells obtained using high throughput RNA-sequencing technique have been used as the experimental data to evaluate the effectiveness of the appropriate step implementation. Applying of the proposed technique involved removing low expressed genes at the first step. The number of genes was reduced from 53186 to 7435 at this step. The filtered gene expression profiles normalizing was considered at the next step. The quality of data normalizing was evaluated by both various graphic tools and using quantitative criterion, which was calculated based on the cluster analysis for the samples which were previously distributed into clusters.

Keywords: RNA sequencing analysis, gene expression profiles, Bioconductor, reducing, normalizing, exploratory analysis, clustering, clustering quality criterion

1 Introduction

Development of technique of gene expression data processing in order to allocate high expressed genes, which allows us to distinguish investigated samples, is one of the current directions of modern bioinformatics. Implementation of this technique create the conditions to reconstruct gene regulatory networks with high level of sensitivity. The further simulation of the reconstructed gene regulatory networks can allow us to better understand the character of genes interconnection and, as a result, it can help us also to understand the ways of influence to the appropriate key genes in order to change the expressions of genes of the network according to goal of the current task. Two techniques are actual to form the array of gene expression

2019 IDDM Workshops.

nowadays: DNA-microchip technique [1,2] and RNA-molecules sequencing method [3,4]. DNA-microchip technique is significantly cheaper in comparison with RNA-molecules sequencing method. As a result of this technique applying, we obtain the matrix of light intensities. Then, four stages should be implemented to transform the light intensities matrix into matrix of genes expression: background correction, normalization, PM correction and summarization. Each of the stage can be implemented by different ways [5–7]. This fact decreases the quality of obtained matrix of genes expression.

Applying RNA-molecules sequencing method allows obtaining the number of investigated genes for studied samples directly. In this reason, this method more exact in comparison with DNA-microchip technique. The number of genes determines the level of this gene activity or its expression. At the next step it is necessary to remove non-expressed genes for all samples and gene with low level of expression. At this stage it is appear the problem identification of boundary value which allows dividing genes to lowly-expressed and highly-expressed. Moreover, the matrix of counts of genes is not suitable for the following processing. Thus, initially, the data should be normalized. This step assumes transformation the counts values into the same suitable range. There are various normalized methods to process gene expression values. However, it should be noted, that the task of objective selection of appropriate normalizing method based on the quantitative criteria has not effective solution nowadays. This fact indicates the actuality of the research.

2 Formal Problem Statement

A block chart of procedure to process the experimental data which are obtained by RNA-molecules sequencing technique is presented in fig. 1. The studied dataset is presented as a matrix of counts, values of which are determined the number of genes corresponding to appropriate sample.

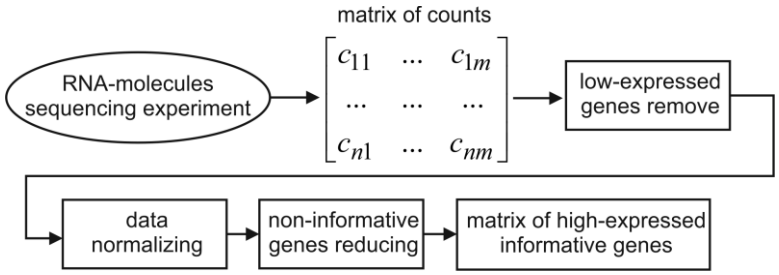

Fig. 1. A step-by-step procedure to transform a matrix of counts to the matrix of highly-expressed informative genes

One of the most important steps of this procedure implementation is the data normalizing. The normalized values of gene expression profiles should have the equal ranges and their norms should be distinguish minimally between each other.

Moreover, the values of genes expression should allow us to identified the samples which belong to various clusters. Considering hereinbefore, evaluation of quality of gene expression profiles normalizing will be performed visually by analysis of both box plot and kernel density plot, and based on quantitative criterion.

The main problem, which is solved within the framework of the current research, consists in comparison analysis of various normalization techniques of gene expression values using different normalizing quality criteria.

3 Literature Review

Various techniques have been proposed over last years to pre-process the results of RNA-molecules sequencing experiments [8-16]. These tools are differed between each other by types and thresholds which are used to process the counts of genes and by algorithms which are used for filtering and alignment of the investigated values. It should be noted that the choice of the alignment algorithm has very great influence to the evaluation accuracy of RNA molecules abundance in the sequenced samples. In this reason, testing different tools for processing of data of RNA-molecules sequencing experiments can help us to choose the best technique for current type of data.

After implementation of the alignment procedure, it is necessary to normalize the recovered values of miRNA counts for purpose of removing variations in the data which have not biological origins and, as a result, can influence to the ranges of measured values change. Correct applying the normalizing technique allows minimizing the experimental and the technical bias without noise introduce. In [17,18] the authors have proposed several normalizing techniques for data of RNA-molecules sequencing experiments. As a result of comparison of different normalization methods effectiveness, the conflicting results were obtained in these works. So, the authors in [18] proposed using the locally weighted linear regression and quantile normalizing techniques. At the same time, they were discouraging against the use of trimmed mean of M values (TMM technique). The obtained results were validated based on the use of polymerase chain reaction (qPCR). In [17] the authors proposed the opposite, to use against quantile normalization the TMM method. The simulation results were used to confirm these findings. An assessment of the relative effectiveness of various pre-processing techniques in terms of statistical criteria, bias, sensitivity and specificity in order to detect the differential expressed genes can be achieved on the basis of complex implementation of both qualitative and quantitative normalizing quality criteria using current techniques of data processing [19–26].

The aim of the paper is exploratory analysis of various technique of data from RNA-molecules sequencing experiment processing based on the complex use of Bioconductor package tools and various quality criteria to estimate the effectiveness of the data processing.

4 Materials and Methods

4.1 Data Set

We used the dataset GSE129336 generated from Gene Expression Omnibus (GEO) database [27] as the experimental data during the simulation process. The data contains the results of expression profiling by high throughput sequencing in human SH-SY5Y neuroblastoma cells [28]. The transcriptomic responses to Mn dose (0,1,5,10,50,100 μM MnCl2 for 5 h) in the investigated cells with three biological replicates per Mn treatment were examined during the experiment performing. Thus, the examined samples can be divided into six clusters considering the Mn dose. Each of the clusters contains three samples. This fact can be used to calculate one of the criteria to estimate the quality of gene expression values processing. Each of the samples contained 53186 of genes. So, the initial dataset contained 53186 of genes or rows and 18 of columns or samples. The early analysis has shown, that there were 27838 non-expressed genes (zero for all samples). Of course, these genes can be removed from the data at the first step. Moreover, the lowly-expressed genes for all samples can be removed from the data too. The search of the thresholding value to remove lowly-expressed genes is one of the solved tasks within the framework of this research.

4.2 Removing Lowly-expressed Genes

As was noted in the section 4.1, the studied dataset contains 53186 of genes. However, 27838 of genes are non-expressed for all samples (the count value is zero). Thus, the number of the expressed genes can be changed from 53186 to 25348 of genes.

At the next step, it is necessary to remove lowly expressed genes considering the appropriate thresholding value. The initial values of the counts of genes are not suitable for solve this task since the range of the genes count value change is very large (in the case of our dataset this range is changed from o to 47434890). In this case it is necessary to transform the count value scale into other, more suitable scale. To solve this tack, Bioconductor package contains *cpm()* function which allows transforming the counts values into count-per-million values as follows:

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \cdot 10^6 \quad (1)$$

where n is number of rows, x_{ij} is value in i-th row and j-th column. Applying this function allows us to obtain the new, more suitable, range of the data values change (from 0 to 380367.8).

The main idea for lowly-expressed genes removing is the following: the use a nominal thresholding of 1 *cpm* value (this value is corresponded to 0 *log2(cpm)* value) allows dividing the genes into two groups (expressed and unexpressed). If value of

gene expression is more than this threshold, the gene is identified as expressed. Otherwise, the gene is identified as unexpressed. Considering the number of samples in the clusters we can suppose that the genes should be expressed in at least one cluster (three samples) for the further analysis.

4.3 Normalizing Gene Expression Profiles

The following normalizing techniques were evaluated during the simulation process: 1) lcpm; 2) TMM; 3) TMMwsp; 4) RLE; 5) upper quartile scaling. Brief describing each of these techniques is presented below.

1. lcpm – the simplest normalizing technique, the log2(cpm) values are calculated during this technique implementation.

2. TMM – trimmed mean of M is the normalizing technique by total count of scaling. The counts quantity for an appropriate target for all samples is estimated during TMM technique implementation. If an expression value is identified in the same proportion for all samples, this gene is identified as non-differentially expressed. It should be noted that this technique does not allow considering the potentially different RNA molecules which are presented in the samples. Applying this method allows us to calculate a linear scaling index for appropriate sample considering weighted average after transforming the data using log fold-changes (M) relative to the absolute intensity in the reference sample (A) [29].

3. TMMwsp – TMM with singleton pairing. This technique is a variant of TMM, in which the data with a high proportion of zeros are processed. Implementation of the TMM method assumes that the genes which have zero value in either library are ignored when pairs of libraries are compared between each other. As opposed to TMM method, implementation of the TMMwsp technique assumes that the positive counts from such genes are reused to increase the quantity of features which are used to compare the libraries. The singleton positive counts are paired up between the libraries in decreasing order of size and then a slightly modified TMM method is applied to the re-ordered libraries.

4. RLE – relative log expression technique. Implementation of this method assumes that the median library is calculated from the geometric average of all columns and the median ratio of each sample to the median library is used as the scale factor.

5. Upper quartile scaling – is the upper-quartile normalizing technique, in which the scale factors are calculated from the 75% quantile of the counts for each of the libraries, after removing genes that are zero in all libraries.

4.4 Quantitative Criterion to Estimate the Quality of Data Normalizing

The main idea to evaluate the quality of data normalizing is the following: as we noted hereinbefore, the samples can be divided into six clusters considering the dose of Mn. Each of the clusters in this case contains three samples. It is naturally that informativity of gene expression profiles is determined by their ability to distinct the samples in different clusters. Thus, the quality of data normalizing can be estimated

based on clustering quality criterion which should consider the samples distribution within clusters and the clusters distribution in the feature space. Considering the high dimension of the studied vectors, the correlation metric should be used to estimate the proximity level between the vectors. This quality criterion of the samples and clusters grouping was calculated as multiplicative combination of Calinski-Harabasz criterion and WB-index [30,31]:

$$QNC = \frac{N_c(N_c - 1) \cdot QCW^2}{(N_s - N_c) \cdot QCB^2} \tag{2}$$

where: N_c is the clusters quantity; N_s is the samples quantity; QCW is an average distance from samples to centers of the clusters where these samples are allocated; QCB is an average distance between clusters' centers. It should be noted that minimum value of the criterion (2) corresponds the best normalizing technique.

5 Experiments, Results and Discussions

Fig. 2 presents the results of lowly expressed genes reducing in accordance with technique which are described in 4.2. To increase the charts informativity the data preliminarily were transformed using *log2(cpm)* function. The number of genes was reduced at this step from 25348 to 7435. The analysis of the obtained in fig. 2 diagrams allows concluding that level of genes expression informativity significantly increased due to remove lowly expressed values. The same conclusion can be done based on the box plots analysis (See Fig. 3). In the case of filtered data use the values of gene expressions for all samples are distributed more evenly and they are shifted to the side of larger values.

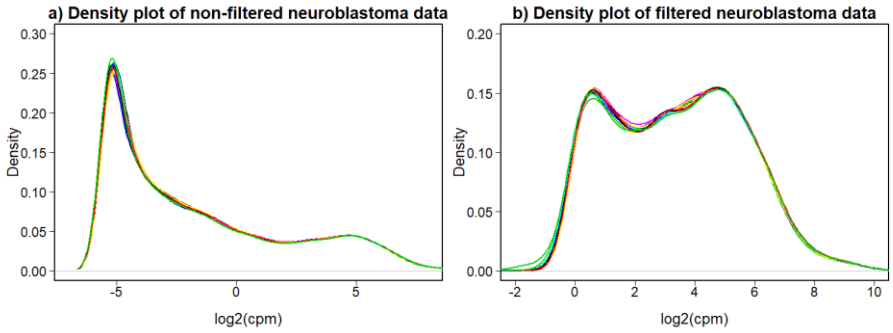

Fig. 2. Density plots of non-filtered and filtered gene expression values distribution for neuroblastoma data samples

The next step of the data preprocessing is their normalizing. Fig. 4 shows the chart of the clustering quality criterion (2) versus the normalizing method. To calculate this criterion values the data previously were divided into clusters considering the Mn

dose. It should be noted that in the case of non-normalized filtered data the value of this criterion was 100,05.

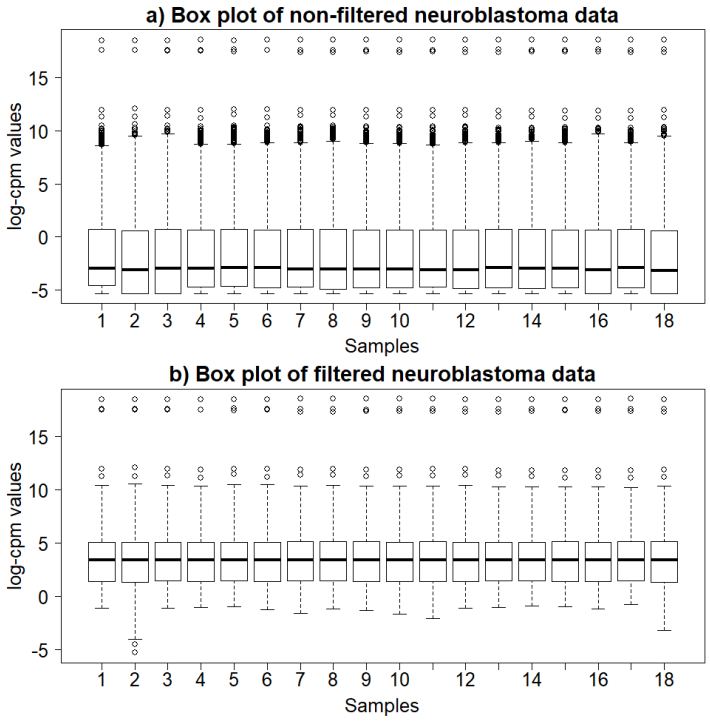

Fig. 3. Box plots of non-filtered and filtered gene expression values distribution for neuroblastoma data samples

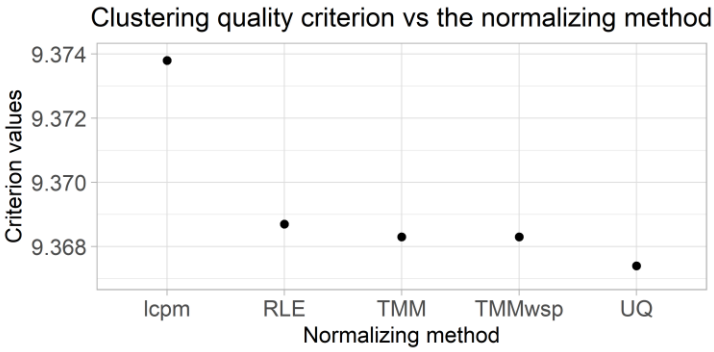

Fig. 4. Dot plot of the quality criterion versus the normalizing method

The obtained results analysis allows concluding that normalizing process significantly increases the quality of the data in terms of the quality criterion (2). The

value of this criterion for non-normalizing data 100,05 decreases more than 10 time. Comparison analysis of various normalizing methods has shown that the easiest *lcpm* method is showed the worst results in comparison with other methods. The difference between methods *TMM, TMMwsp, RLE* and *Upper quartile scaling* is very small, however, the value of the criterion (2) achieved the minimum one in the case of Upper quartile scaling method apply. This fact indicates the reasonable of this method use for current type of data normalizing.

At the next step it is necessary to remove heteroscedasticity from the data. The analysis of the normalized data has shown that in the case of RNA-seq data use, the variance values are not depend on the mean values. Methods that counts of the model with the use of Negative Binomial distribution are based on a quadratic mean-variance relationship. In *limma* package of R software, linear modelling is performed using the normalized values. In this case the data should be normally distributed and the mean-variance relationship is evaluated with the use of precision weights calculated by the *voom()* function. Fig. 5 presents the results of this step implementation.

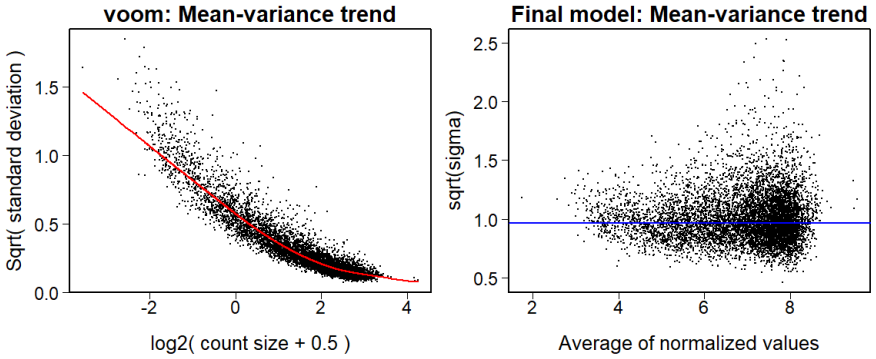

Fig. 5. Visualization of the heteroscedasticity removing from the data

The left chart in the Fig.5 shows the mean-variance relationship of normalized gene expression values. Usually, the voom-plot shows a decreasing trend between the means and variances which are appeared due to an existence of both the technical incorrectness during the sequencing experiment performing and the biological variation among the replicate samples from various cell samples. Typically, the results of the experiments with high level of biological variation are presented as a flatter trends. The variance values in this case are not significantly changed for high expression values (right chart in Fig. 5). And otherwise, experiments including data with low biological variation usually have tend to sharp decreasing the variance values. Moreover, the voom-plot allows us to visual evaluate the quality of gene expression filtration process. If filtration process of lowly-expressed genes is insufficient, then, the variance values should be decreased at the low end of the expression scale due to very small gene expression values.

In order to visual summarize the results for all genes in obtained groups, we create a mean-difference plots using the *plotMD* function of *limma* package. These plots allow us to display log-Fold-change values from the linear model which can be fitted against the average of log-expression values. These charts allow us to identify differentially expressed genes. The charts are shown in Fig. 6.

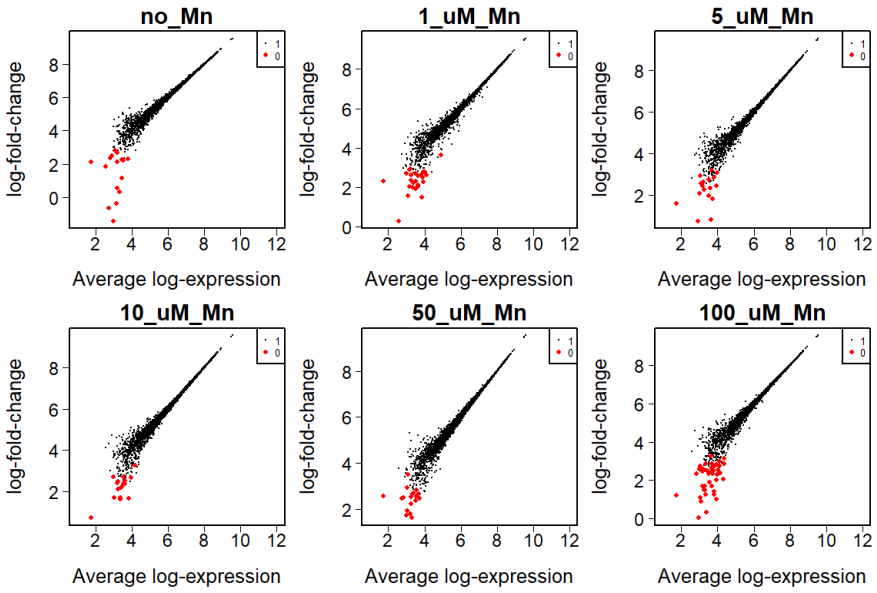

Fig. 6. Mean-difference plots of gene expression profiles for investigated groups

Result of the visual analysis of the obtained diagrams allows concluding the greatest number of genes in investigated groups have high level of differentially expression (black color or number 1). It means that these genes are informative to distinct the samples for the further processing. However, the data contains some quantity of lowly-expressed genes (red color or zero number). It is means that these data need the following processing for purpose of non-informative genes reducing in terms of various quantitative criteria.

Fig. 7 shows the box charts of the processed gene expression profiles. The samples previously were reordered considering Mn dose from 0 to 100 μM. Analysis of character of gene expressions distribution allows us to conclude about correctness of data preprocessing step implementation. The values of gene expression have the same and not so much ranges, all genes are enough highly-expressed for all of the samples. However, it should be noted, that there is some quantity of lowly-expressed genes (for example, in Mn_1 sample). This fact indicates about necessity the further data processing on the basis of the use of current techniques of complex data processing.

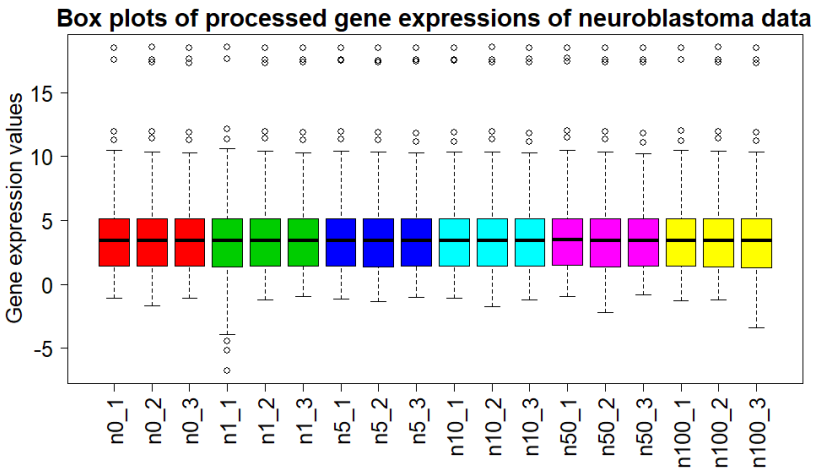

Fig. 7. Results of gene expression profiles of neuroblastoma data processing

6 Conclusions

This paper presents the research results about processing of RNA-molecules sequencing experiments. The dataset GSE129336 generated from Gene Expression Omnibus database was used as the experimental one. This data contains the results of expression profiling by high throughput sequencing in human SH-SY5Y neuroblastoma cells. The initial data matrix contained counts of expressed genes for studied samples. At the first step, we have removed lowly-expressed genes. The number of genes was changed from 53186 to 7435. Then, we have compared various normalizing technique using clustering quality criterion as the main criterion of appropriate normalizing method effectiveness estimation. At the next steps we have analyzed the obtained results using various visualization techniques. The analysis of the processed genes expression values distributions allows concluding about high effectiveness of the proposed technique, since its implementation allows allocating a set of similarly distributed highly-expressed genes for the following processing.

References

1. Cha, Y.J., Park, S.M., You, R., Kim, H., Yoon, D.K.: Microstructure arrays of DNA using topographic control. Nature Communications, 10(1), art. no. 2512 (2019) doi: 10.1038/s41467-019-10540-2
2. Nagashima, M., Miwa, N., Hirasawa, H., Katagiri, Y., Takamatsu, K., Morita, M.: Genome-wide DNA methylation analysis in obese women predicts an epigenetic signature for future endometrial cancer. Scientific Reports, 9(1) (2019), art. no. 6469 doi: 10.1038/s41598-019-42840-4

3. Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I., Wilson, A.C.: Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nature Communications, 10(1), art. no. 754 (2019) doi: 10.1038/s41467-019-08734-9
4. Lian, B., Hu, X., Shao, Z.-M.: Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. Scientific Reports, 9(1), art. no. 6032 (2019) doi: 10.1038/s41598-019-42184-z
5. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19 (2), pp. 185-193 (2003) doi: 10.1093/bioinformatics/19.2.185
6. Affymetrix. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA, pp. 1-27 (2002)
7. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. Selected Works of Terry Speed, pp. 601-616 (2012) doi: 10.1007/978-1-4614-1347-9_15
8. Buermans, H.P.J., Ariyurek, Y., van Ommen, G., den Dunnen, J.T., 'T Hoen, P.A.C.: New methods for next generation sequencing based microRNA expression profiling. BMC Genomics, 11(1), art. no. 716 (2010) doi: 10.1186/1471-2164-11-716
9. Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J.M., Aransay, A.M.: miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Research, 37 (SUPPL. 2), pp. W68-W76 (2009) doi: 10.1093/nar/gkp347
10. Farazi, T.A., Brown, M., Morozov, P., ten Hoeve, J.J., Ben-Dov, I.Z., Hovestadt, V., Hafner, M., Renwick, N., Mihailović, A., Wessels, L.F.A., Tuschl, T.: Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. Methods, 58 (2), pp. 171-187 (2012) doi: 10.1016/j.ymeth.2012.07.020
11. Hackenberg, M., Rodríguez-Ezpeleta, N., Aransay, A.M.: MiRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Research, 39 (SUPPL. 2), pp. W132-W138 (2011) doi: 10.1093/nar/gkr247
12. Huang, P.-J., Liu, Y.-C., Lee, C.-C., Lin, W.-C., Gan, R.R.-C., Lyu, P.-C., Tang, P.: DSAP: Deep-sequencing small RNA analysis pipeline. Nucleic Acids Research, 38 (SUPPL. 2), art. no. gkq392, pp. W385-W391 (2010) doi: 10.1093/nar/gkq392
13. Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C.J., Marra, M.A.: Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research, 18 (4), pp. 610-621 (2008) doi: 10.1101/gr.7179508
14. Pantano, L., Estivill, X., Martí, E.: SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Research, 38 (5), art. no. gkp1127, pp. e34.1-e34.13 (2009) doi: 10.1093/nar/gkp1127
15. Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q., Shen, B.: Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic Acids Research, 40 (10), pp. 4298-4305 (2012) doi: 10.1093/nar/gks043
16. Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., Shomron, N.: miRNAkey: A software for microRNA deep sequencing analysis. Bioinformatics, 26 (20), art. no. btq493, pp. 2615-2616 (2010) doi: 10.1093/bioinformatics/btq493

17. Dillies, M.-A., Rau, A., Aubert, J., at.al.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics, 14 (6), art. no. bbs046, pp. 671-683 (2013) doi: 10.1093/bib/bbs046
18. Garmire, L.X., Subramaniam, S.: Evaluation of normalization methods in mammalian microRNA-Seq data. RNA, 18 (6), pp. 1279-1288 (2012) doi: 10.1261/rna.030916.111
19. Izonin, I., Trostianchyn, A., Duriagina, Z., Tkachenko, R., Tepla, T., Lotoshynska, N.: The combined use of the wiener polynomial and SVM for material classification task in medical implants production. International Journal of Intelligent Systems and Applications, 10 (9), pp. 40-47 (2018) doi: 10.5815/ijisa.2018.09.05
20. Tam, S., Tsao, M.-S., McPherson, J.D.: Optimization of miRNA-seq data preprocessing. Briefings in Bioinformatics, 16 (6), pp. 950-963 (2015) doi: 10.1093/bib/bbv019
21. Burov, Y., Vysotska, V., Kravets, P.: Ontological approach to plot analysis and modeling. CEUR Workshop Proceedings, 2362, (2019)
22. Lytvyn, V., Vysotska, V., Dosyn, D., Burov, Y.: Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology, 15 (2), pp. 66-85 (2018)
23. Hu, Z., Mashtalir, S.V., Tyshchenko, O.K., Stolbovyi, M.I.: Clustering matrix sequences based on the iterative dynamic time deformation procedure. International Journal of Intelligent Systems and Applications, 10 (7), pp. 66-73 (2018) doi: 10.5815/ijisa.2018.07.07
24. Hu, Z., Bodyanskiy, Y.V., Tyshchenko, O.K., Tkachov, V.M.: Fuzzy clustering data arrays with omitted observations. International Journal of Intelligent Systems and Applications, 9 (6), pp. 24-32 (2017) doi: 10.5815/ijisa.2017.06.03
25. Babichev, S., Škvor, J., Fišer, J., Lytvynenko, V.: Technology of gene expression profiles filtering based on wavelet analysis. International Journal of Intelligent Systems and Applications, 10 (4), pp. 1-7 (2018) doi: 10.5815/ijisa.2018.04.01
26. Babichev, S.A., Gozhyj, A., Kornelyuk, A.I., Lytvynenko, V.I.: Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm. Biopolymers and Cell, 33 (5), pp. 379-392 (2017) doi: 10.7124/bc.000961
27. NCBI Homepage, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129336
28. Fernandes, J., Chandler, J.D., Lili, L.N., Uppal, K., Hu, X., Hao, L., Go, Y.-M., Jones, D.P.: Transcriptome analysis reveals distinct responses to physiologic versus toxic manganese exposure in human neuroblastoma cells. Frontiers in Genetics, 10 (JUN), art. no. 676 (2019) doi: 10.3389/fgene.2019.00676
29. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), art. no. r25 (2010) doi: 10.1186/gb-2010-11-3-r25
30. Babichev, S., Krejci, J., Bicanek, J., Lytvynenko, V.: Gene expression sequences clustering based on the internal and external clustering quality criteria. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, art. no. 8098744, pp. 91-94 (2017) doi: 10.1109/STC-CSIT.2017.8098744
31. Babichev, S.: An evaluation of the information technology of gene expression profiles processing stability for different levels of noise components. Data, 3 (4), art. no. 48 (2018) doi: 10.3390/data3040048