Open Journal of Statistics, 2016, 6, 628-636
Published Online August 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.64053

Challenges Analyzing RNA-Seq Gene Expression Data

Liliana López-Kleine, Cristian González-Prieto

Department of Statistics, Universidad Nacional de Colombia—Sede Bogotá, Bogotá, Colombia
Email: llopezk@unal.edu.co

Received 25 June 2016; accepted 16 August 2016; published 19 August 2016

Abstract

The analysis of messenger Ribonucleic acid obtained through sequencing techniques (RNA-sequencing) data is very challenging. Once technical difficulties have been sorted, an important choice has to be made during pre-processing: Two different paths can be chosen: Transform RNA-sequencing count data to a continuous variable or continue to work with count data. For each data type, analysis tools have been developed and seem appropriate at first sight, but a deeper analysis of data distribution and structure, are a discussion worth. In this review, open questions regarding RNA-sequencing data nature are discussed and highlighted, indicating important future research topics in statistics that should be addressed for a better analysis of already available and new appearing gene expression data. Moreover, a comparative analysis of RNAseq count and transformed data is presented. This comparison indicates that transforming RNA-seq count data seems appropriate, at least for differential expression detection.

Keywords

RNA-Seq Analysis, Count Data, Preprocessing, Differential Expression, Gene Co-Expression Network

1. Introduction

This sequencing of messenger RNA transcripts (RNA-seq) is a recently developed approach to gene expression or transcriptome profiling that uses deep-sequencing technologies. Studies using this method have allowed assessing the complexity of transcriptomes. RNA-seq also provides more precise measurement of levels of transcripts and their isoforms than other methods based on hybridization (such as microarrays), that were used previously, but poses also new challenges [1]. Great issues concerning the identification of the real number of RNA fragments taking into account isoforms, mitochondrial and ribosomal RNA have appear but are beyond the interest of this review. Several satisfactory developments assure a good characterization of RNA-seq transcripts

How to cite this paper: López-Kleine, L. and González-Prieto, C. (2016) Challenges Analyzing RNA-Seq Gene Expression Data. *Open Journal of Statistics*, **6**, 628-636. http://dx.doi.org/10.4236/ojs.2016.64053

[2] to be used for increasing comprehension of biological knowledge. Here, statistical challenges that arise once RNA counts are obtained (after mapping), are discussed.

1.1. RNA-Seq Statistical Challenges

During the last 15 years, statistical research has been done, driven by the need to analyze properly data from high-throughput genomic assays, in particular microarrays. In the last five years, high-throughput sequencing technology has been changing the face of biological research, replacing the old microarray technology. As mentioned by Datta and Net-tleton (2014), with any new high-throughput technology come new data analytic challenges that have been solved in proposing new analytical methods based on novel and older concepts of error rate control for testing multiple hypotheses, various adaptations of existing [3].

Analyzing mapped reads is a major challenge than continuous microarray data, because count data has to be modeled using discrete distributions that had not been used so far for gene expression data analysis. Moreover, an issue concerning dimensionality appears, because often less replicate samples are available than were for microarray data. Even though, data produced using these technologies are proving to be the most informative of any thus far, very little attention has been paid to fundamental design aspects of data collection and analysis, namely sampling, randomization, replication, and blocking [4] (Auer and Doerge, 2010).

RNA-seq data and its proper analysis has an enormous potential to promote genomic research and enhance understanding of biological processes, but a detailed comprehension of this technology and the type of data produced is needed in order to obtain confident results. In this review open questions regarding RNA-sequencing data nature are discussed and highlighted, indicating important future research topics that should be addressed for a better analysis of already available and new appearing gene expression data. A comparative analysis of RNAseq count and transformed data is also presented allowing interesting results that make count data transformations generally applicable.

1.2. RNA-Seq Data Preprocessing

As for microarray data, several similar steps of preprocessing need to be achieved before RNA-seq data can be used for analysis. Nevertheless, two main paths can be chosen for RNA-seq data. The first one is to transform the count data to a continuous variable using RPKM (reads per kilobase per million mapped reads) as originally introduced by [5] and the second path is to continue statistical analysis with count data as it is. Each path requires different analytic tools because each type of data need to be treated in a different way.

1.2.1. Transformation of RNA-Seq Count Data into a Continuous Variable

In the case of transformation to RPKM, the preprocessing begins by equalizing sequencing depths, to compare the ex-pression measures across different genes and samples. These “normalization” is made by dividing counts by gene length (a variable) and the total amount of reads in each experiment (a constant). Then, analysis conceived for continuous microarray data are applied without apparent concern about the distribution differences between these transformations on count data compared to transformations done on continuous microarray data. Several authors have discussed inconsistencies but no deep discussion on distribution of RPKM has been done [6]-[8].

More realistic models than RPKM addressed the case for multiple isoforms [9] proposing a Poisson distribution for counts and create a continuous variable during the mapping process. The major assumption was that the number of reads coming from an exon of a certain length is Poisson where the mean is a normalized function of the exon length. The first insert length model extended the approach of [9] to paired-end reads [10]. Their algorithm was made available through the software called Cuffllinks in early 2011 and uses FPKM (fragments instead of reads). FPKM is based on a probabilistic assignment method indicating the probability that a fragment selected at random originates from a given transcript [11]. Similarities between both types of estimations have been reported, but again no known discussion on data distribution of FPKM has been undertaken. FPKM is analyzed with Cufflinks developed especially for this data transformation of RNA-seq counts. Therefore, use of FPKM seems a more restricted transformation than RPKM.

1.2.2. Preprocessing or RNA-Seq Count Data without Transformation

Raw read counts from different experiments are not directly comparable without adjustment for technical variation due to sequencing depth (a process also called normalization). Complex normalization schemes for RNA-seq data have been proposed by [12]-[14] Bullard *et al.*, 2010 Robinson and Oshlack, 2010. Sample specific

normalizations are combined with library sizes in these methods. Trimmed mean of M-values normalization (TMM) [14] and the normalization scheme provided by [13] are among the most efficient and easy to use. When these path is chosen, new methods developed for counts are used for analysis.

1.3. Maintaining the Integrity of the Specifications

Depending on the type of normalization or transformation that has been undertaken to raw RNA-seq data, several tools are available for subsequent analysis.

1.3.1. Differential Expression for RPKM and FPKM

Generally, methods developed for continuous microarray data are applied on RPKM [5] transformed data. For FPKM, [11] have developed the algorithm Cuffdiff for differential expression. It estimates the expression transcript-level resolution and controls for variability across replicates. Following the number of citations of these two articles during 2016 (google academics consulted on march 20th 2016), both types of transformations are almost equally used (157 citations for [5] and 140 for [11]). No methods developed based on distributional properties have been proposed but are also rare for microarray data [15].

1.3.2. Count Data

Several statistical methods and related R packages for differential gene expression analysis based on RNA-seq data have been developed over the years. The packages DESeq [13] and EdgeR [16] are a popular choice amongst users of RNA-seq. BaySeq [17] is a Bioconductor package that identifies differential expression using high throughput sequencing data via empirical Bayesian methods. Another method, called TSPM, is based on a two-stage Poisson model [18]. These four methods are compared by [19]. The results suggest that baySeq performs best in terms of ranking genes according to their significance to be declared differentially expressed. Both edgeR and DESeq perform similarly and close to baySeq. The results from TSPM are most variable and often the poorest when the number of replicates is small [19].

One year later, six methods where compared by [20]: DESeq, DEGseq, edgeR, NBPSeq, TSPM and baySeq using both real and simulated data with the result that all six methods produce similar fold changes and reasonable overlapping of differentially expressed genes based on p-value, edgeR being little bit superior. However, all six methods suffer from over-sensitivity as reported by the authors.

A recent and not yet popular method based on a hierarchical negative binomial model (borrowing information across gene-variety means and across gene-specific over dispersion parameters) and using a computationally tractable empirical Bayes approach to inference has been proposed by [20].

Also, threshold-independent methods for particular cases, for example the detection of marginal expression changes in cognitively stratified patients at different disease stages, have been developed [21]. This approach is based on the comparison of the distribution of changes in a well-defined gene group with the global distribution of the experiment [21].

The question of preprocessing and making samples comparable is not yet completely solved for microarrays and very far from being solved for RNA-seq data. RPKM seems like a general solution for prokaryotes because the nature of transformed data allows using methods developed for microarrays, but limitations have been shown, especially for eukaryotes [10]. Algorithms for all other particular cases are being developed. Although RNA-seq data has been analyzed and biological conclusions have been drawn, a main question is still unanswered: How general is RPKM, can it really be analyzed with algorithms developed for microarray data? An answer to this question, based on empirical observations, is given in Section 4.

A second open question is: Can we combine microarray data and RNA-seq data to increase biological knowledge or do we need to repeat all microarray experiments? Assess this comparison is tricky because influence of experiment type, data type and method would confounded. Nevertheless, even a partial answer to this question would be of general interest for biologists and applied statisticians.

2. Comparison of Differential Gene Expression between RPKM and Count Data

2.1. Empirical Characterization of RPKM Data Distribution

2.1.1. Simulation of Count Data

The function make Example Count Data Set from the DeSeq Bioconductor package [13] was used to simulate

ten RNAseq count tables with 20,000 genes (rows) and 100 samples (50 treatments and 50 controls) as follows:

1) Mean expression values were sampled from an exponential distribution with parameter 1/250.

2) Once graphically verified, one table of count data with 50 treatments and 50 controls, and a proportion of differentially expressed genes of 0.3 was constructed.

3) Mean values of gene expression are divided in two conditions (treatment and control) such that the log2 fold-change is still centered in 0 and with a standard deviation of 2.

4) Counts were sampled from a negative binomial distribution with mean values of gene expression mentioned before and multiplied by size factors between 0.1 and 2.55. The size parameter of the binomial distribution was set at 1/0.2.

2.1.2. Simulations of Gene Length

Using the real gene length of three organisms obtained from NCBI (Escherichia coli, Homo sapiens and Arabidobsis thaliana) the package fitdistrplus [22] was used to adjust a distribution to these data. We concluded that the gene length distribution is gamma with outliers.

In order to simulate gene lengths of the 20.000 genes present in the count data table, the mean of the gamma distribution was estimated based on mean gene length of the three organisms we had taken as example. Thousand extreme values were sampled from a distribution with different mean (third quartile of 19.000 simulated gene lengths) to end up with 20.000 gene lengths.

2.1.3. RPKM Transformed Data

One count tables and gene lengths were generated as explained above. Those counts were then transformed to continuous RPKM data [5] dividing each count value by the simulated gene length and the total sum of counts for each column (sample). Finally, RPKM data were log2 transformed. The result of the density adjusted to 100 samples of one of the simulations (treatments and controls) can be observed in **Figure 1**.

2.1.4. Verification of RPKM Distribution

With the estimated mean and variance parameters of each of the 100 simulated samples (50 treatments, 50 controls with 0.3 proportion of differentially expressed genes), 100 samples of normal distribution were generated and adjustment was tested using Kolmogorov-Smirnov test. As shown in **Figure 2**, most p-values are not significant, indicating that the transformed RPKM data can be considered having a normal distribution.

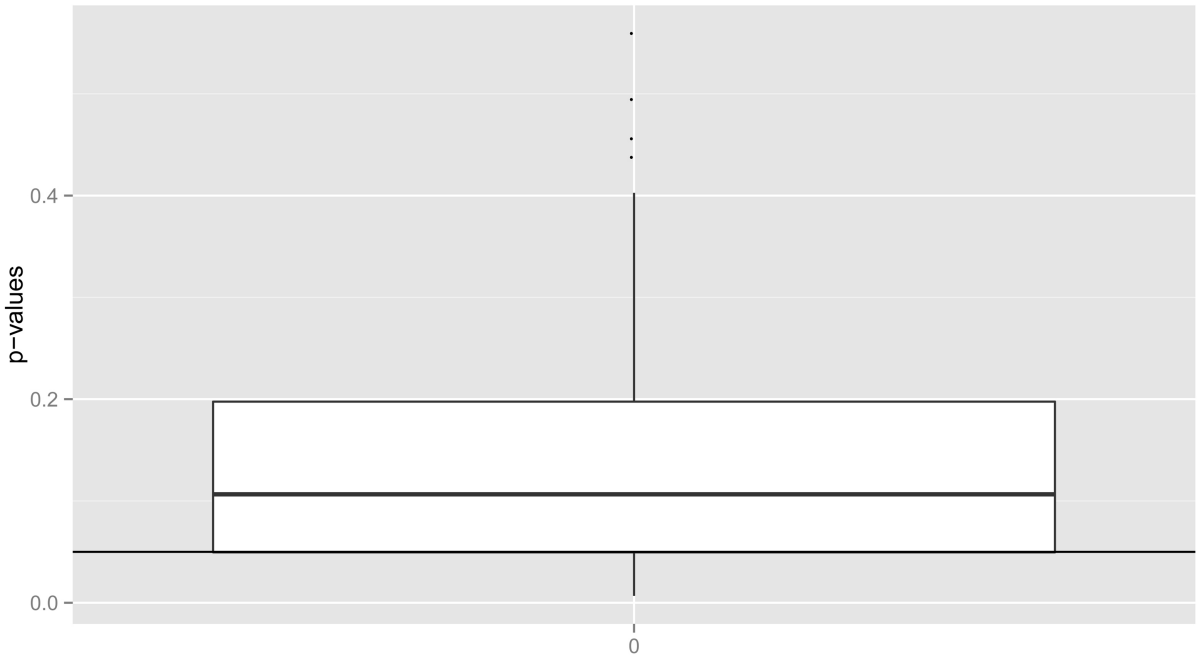

Figure 1. Densities of 100 simulated samples of RPKM transformed count data.

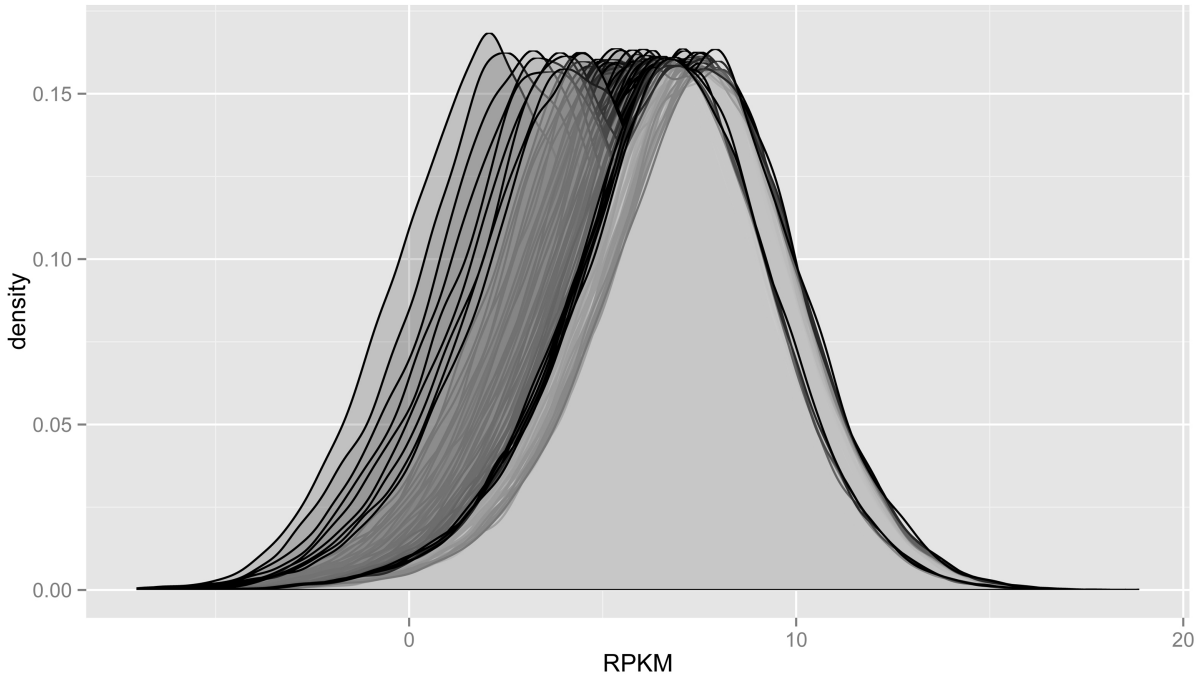

Figure 2. Boxplot of 100 p-values obtained from Kolmogorov-Smirnov goodness to fit test comparing transformed RPKM data to a theoretical normal distribution. Red line indicates a p-value of 0.05.

2.2. Analysis of RNA-Seq Count Data without Transformation

2.2.1. Real Data

Three available data sets (NCBI, Genome Expression Omnibus) comparing two conditions (controls vs. treatments) were analyzed using DeSeq standard normalization and detection of differentially expressed genes [13]:

- GSE67402: Controlled measurement and comparative analysis of cellular components in *E. coli* reveals broad regulatory changes under long-term starvation [23].
- GSE76268: Integration of ATAC-seq and RNA-seq Identies Human Alpha Cell and Beta Cell Signature Genes [24].
- GSE72548: RNA-seq analysis of Arabidopsis thaliana wild-type roots and type-A arr3, 4, 5, 6, 7, 8, 9, 15 mutant roots non-infected and infected with Heterodera schachtii nematodes [25]. Results are shown in **Table 1**.

2.2.2. Simulated Data

Deseq results on 10 simulated count data tables are shown in **Table 2** and indicate (as expected) approximately 30% of genes with differential expression.

2.3. Analysis of RPKM Transformed Data

2.3.1. Real Data

Count tables of the three experiments used above, were RPKM and log2 transformed to be analyzed using Significance Analysis of Microarray (SAM) [26], a standard method for microarray data analysis. Moreover, we retained the propor-tion of genes common to both analysis (**Table 3** and **Table 4**). The results indicate that the proportion and identity of genes identified on count data using Deseq or on RPKM transformed data are equivalent.

2.3.2. Conclusion on RPKM vs. Count Data Comparison

The most important conclusion of the above comparison is that RPKM transformation with posterior log2 normalization conduces to a data distribution which is very similar to continuous microarray data and that tools

Table 1. Table indicating number of differentially expressed genes (DEG), proportion of the total (POT) using Deseq on real data. FDR for these results is lower than 0.0001. COUNT-RPKM indicates the proportion of genes that were detected by both methods.

Experiment	DEG	POT	COUNT-RPKM
GSE67402	776	0.173	0.834
GSE76268	1350	0.067	0.951
GSE72548	9955	0.296	0.798

Table 2. Table indicating number of differentially expressed genes (DEG), proportion of the total (POT) Deseq on simulations. FDR for these results is lower than 0.0001. COUNT-RPKM indicates the proportion of genes that were detected by both methods.

Simulation	DEG	POT	COUNT-RPKM
1	5458	0.273	0.972
2	5969	0.298	0.903
3	5952	0.297	0.902
4	6002	0.300	0.885
5	5714	0.286	0.917
6	6323	0.316	0.850
7	5486	0.274	0.966
8	6011	0.301	0.896
9	5923	0.296	0.902
10	5889	0.294	0.892

Table 3. Table indicating number of differentially expressed genes (DEG) and proportion of the total (POT) in real data sets using SAM. FDR for these results is lower than 0.0001.

Experiment	DEG	POT
GSE67402	834	0.186
GSE76268	967	0.048
GSE72548	10,796	0.321

Table 4. Table indicating number of differentially expressed genes (DEG), proportion of the total (POT) in simulations using SAM. FDR for these results is lower than 0.0001.

Simulation	DEG	POT
1	5306	0.265
2	5392	0.269
3	5366	0.268
4	5311	0.266
5	5242	0.262
6	5378	0.2689
7	5302	0.265
8	5383	0.269
9	5342	0.267
10	5253	0.263

developed for the analysis of them can be used and will conduct to almost similar results. Moreover, the analysis show that this statement is true: Results obtained using methods developed for count data were very similar to results obtained when count data was RPKM transformed and tools developed for continuous microarray data are used. Therefore, it is safe to conclude that RPKM transformation conducts to a similar normalization and that analysis tools developed for microarrays can be used. This also is encouraging regarding the combined analysis of RNA-seq and microarray data. Nevertheless, caution is still required, until an analytic characterization of RPKM transformation is done to confirm the here presented results.

2.3.3. Tools for RNA-Seq Co-Expression Networks

Reconstructing gene or protein networks is a very important tool in deciphering molecular mechanisms. One of the most important data source for this reconstruction has been gene expression data because it reflects coordinated activity of different genes at the same time. Only few examples of gene reconstruction based on RNA-seq data exist. As for assess-ing differential expression, two separate pathways have to be taken, depending if counts or transformed data is used. Even less discussion on this subject than for detection of differential expression is found in literature. Additionally to the already mentioned challenges, difficulties with the gene profile similarity estimation appear, which should be a measure suitable for count data if counts are not transformed.

Some studies addressed the question of comparing gene co-expression network reconstruction with RNA-seq data, applying Pearson correlation to both types of data avoiding discussion on usefulness of this similarity measure. Iancu *et al.* [27] conducted a study comparing co-expression networks constructed with count data to networks constructed with microarray data using the same method for different data types. They concluded that the RNA-seq coexpression network displayed overlapping structure with the microarray network. Pearson correlations from RNA-seq data were higher and therefore, higher network connectivity, heterogeneity and centrality was observed in the RNA-seq network. A more recent study constructs co-expression networks using also Pearson correlation on a huge Arabidopsis thaliana data set [22]. The authors observe sensitivity to variance stabilizing transformations on RNA-seq data but overall similarities between RNAseq networks and microarray networks.

Gene co-expression construction is a complex procedure and still open questions exist when used on microarray data [28]. These difficulties need to be addressed for RNA-seq data as well, but the most important open question regarding RNA-seq networks, is: What similarity measure is more appropriate. Is it right to apply Pearson correlation? Or is it better to use similarity measures for count data, perhaps test and adapt association measures like has been done for other type of data [29]? How do mutual information and other non-linear similarity measures behave? A research in this sense would also shed some light on gene classification and clustering, which is performed on similarity or distance matrices between genes and widely used for annotation purposes and functional prediction.

3. Conclusion

Although, several studies have used and analyzed RNA-seq data with traditional or new developed statistical methods, open questions on data type and nature remain still unanswered and should receive more attention. Throughout this review we have highlighted some open questions and addressed one of them regarding the RPKM tranformation of count data. Adressing all statistical issues related to RNA-seq data analysis is especially important in order to achieve confident results for assessing biological knowledge extracted from gene expression data, which has proven so far to be highly informative.

Ethical Statement

The authors declare that they do not have any conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics*, **10**, 57-63. http://dx.doi.org/10.1038/nrg2484

[2] Ozsolak, F. and Milos, P.M. (2011) RNA Sequencing: Advances, Challenges and Opportunities. *Nature Reviews Ge-*

netics, **12**, 87-98. http://dx.doi.org/10.1038/nrg2934

[3] Datta, S. and Nettleton, D., Eds. (2014) Statistical Analysis of Next Generation Sequencing Data. Springer, New York. http://dx.doi.org/10.1007/978-3-319-07212-8

[4] Auer, P.L. and Doerge, R.W. (2010) Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, **185**, 405-416. http://dx.doi.org/10.1534/genetics.110.114983

[5] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and Quantifying Mammalian tran-Scriptomes by RNA-Seq. *Nature Methods*, **5**, 621-628. http://dx.doi.org/10.1038/nmeth.1226

[6] Kharchenko, P.V., Xi, R. and Park, P.J. (2011) Evidence for Dosage Compensation between X and Autosomes in Mammals. *Nature Genetics*, **43**, 1167-1169. http://dx.doi.org/10.1038/ng.991

[7] Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure is Inconsistent among Samples. *Theory in Biosciences*, **131**, 281-285. http://dx.doi.org/10.1007/s12064-012-0162-3

[8] Wang, L., Wang, S. and Li, W. (2012) RSeQC: Quality Control of RNA-Seq Experiments. *Bioinformatics*, **28**, 2184-2185. http://dx.doi.org/10.1093/bioinformatics/bts356

[9] Jiang, H. and Wong, W.H. (2009) Statistical Inferences for Isoform Expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032. http://dx.doi.org/10.1093/bioinformatics/btp113

[10] Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) Tophat: Discovering Splice Junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-1111. http://dx.doi.org/10.1093/bioinformatics/btp120

[11] Trapnell, C., Roberts, A., Goff, L., Pertea, G, Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cuffflinks. *Nature Protocols*, **7**, 562-578. http://dx.doi.org/10.1038/nprot.2012.016

[12] Bullard, J.H., Purdom, E., Hansen, K.D. and Dudoit, S. (2010) Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics*, **11**, 94. http://dx.doi.org/10.1186/1471-2105-11-94

[13] Anders, S. and Huber, W. (2010) Differential Expression Analysis for Sequence Count Data. *Genome Biology*, **11**, R106. http://dx.doi.org/10.1186/gb-2010-11-10-r106

[14] Robinson, M.D. and Oshlack, A. (2010) A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biology*, **11**, R25. http://dx.doi.org/10.1186/gb-2010-11-3-r25

[15] Tovar, J.R., López-Kleine, L. and Ordoñez, J.A. (2015) Identification of Global Gene Expression Shifts Using Microarray Data from Different Biological Conditions. *Open Journal of Statistics*, **5**, 360-372. http://dx.doi.org/10.4236/ojs.2015.55038

[16] Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics*, **26**, 139-140. http://dx.doi.org/10.1093/bioinformatics/btp616

[17] Hardcastle, T.J. and Kelly, K.A. (2010) BaySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics*, **11**, 422. http://dx.doi.org/10.1186/1471-2105-11-422

[18] Kvam, V.M., Liu, P. and Si, Y. (2012) A Comparison of Statistical Methods for Detecting Differentially Expressed Genes from RNA-Seq Data. *American Journal of Botany*, **99**, 248-256. http://dx.doi.org/10.3732/ajb.1100340

[19] Guo, Y., Li, C.I., Ye, F. and Shyr, Y. (2013) Evaluation of Read Count Based RNAseq Analysis Methods. *BMC Genomics*, **14**, S2. http://dx.doi.org/10.1186/1471-2164-14-S8-S2

[20] Niemi, J., Mittman, E., Landau, W. and Nettleton, D. (2015) Empirical Bayes Analysis of RNA-Seq Data for Detection of Gene Expression Heterosis. *Journal of Agricultural, Biological, and Environmental Statistics*, **20**, 614-628. http://dx.doi.org/10.1007/s13253-015-0230-5

[21] Guffanti, A., Simchovitz, A. and Soreq, H. (2014) Emerging Bioinformatics Approaches for Analysis of NGS-Derived Coding and Non-Coding RNAs in Neurodegenerative Diseases. *Frontiers in Cellular Neuroscience*, **8**, 89. http://dx.doi.org/10.3389/fncel.2014.00089

[22] Giorgi, F.M., Del Fabbro, C. and Licausi, F. (2013) Comparative Study of RNA-Seq- and Microarray Coexpression Net-Works in Arabidopsis Thaliana. *Bioinformatics*, **29**, 717-724. http://dx.doi.org/10.1093/bioinformatics/btt053

[23] Houser, J.R., Barnhart, C., Boutz, D.R., *et al*. (2015) Controlled Measurement and Comparative Analysis of Cellular Components in *E. coli* Reveals Broad Regulatory Changes in Response to Glucose Starvation. *PLoS Computational Biology*, **11**, e1004400. http://dx.doi.org/10.1371/journal.pcbi.1004400

[24] Ackermann, A.M., Wang, Z., Schug, J., Naji, A. and Kaestner, K.H. (2016) Integration of ATAC-Seq and RNA-Seq Identi-Fieses Human Alpha Cell and Beta Cell Signature Genes. *Molecular Metabolism*, **5**, 233-244.

http://dx.doi.org/10.1016/j.molmet.2016.01.002

[25] Shanks, C.M., Rice, J.H., Zubo, Y., Schaller, G.E., Hewezi, T. and Kieber, J.J. (2015) The Role of Cytokinin during Infection of *Arabidopsis thaliana* by the Cyst Nematode *Heterodera schachtii*. *Molecular Plant-Microbe Interactions*, **29**, 57-68. http://dx.doi.org/10.1094/MPMI-07-15-0156-R

[26] Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121. http://dx.doi.org/10.1073/pnas.091062498

[27] Iancu, O.D., Kawane, S., Bottoml, D., Searles, R., Hitzemann, R. and McWeeney, S. (2012) Utilizing RNA-Seq Data for *de Novo* Coexpression Network Inference. *Bioinformatics*, **28**, 1592-1597. http://dx.doi.org/10.1093/bioinformatics/bts245

[28] López-Kleine, L., Leal, L. and López, C. (2013) Biostatistical Approaches for the Reconstruction of Gene Co-Expression Networks Based on Transcriptomic Data. *Briefings in Functional Genomics*, **12**, 457-467. http://dx.doi.org/10.1093/bfgp/elt003

[29] Riaz, M., Munir, S. and Asghar, Z. (2014) On the Performance Evaluation of Different Measures of Association. *Revista Colombiana de Estadística*, **37**, 1-24. http://dx.doi.org/10.15446/rce.v37n1.44353