ECRONICON OPEN ACCESS

EC MICROBIOLOGY
Editorial

Bioinformatics of RNA-Seq based Gene Expression Profiling

Praveen-Kumar Raj-Kumar*

Bioinformatics Scientist at Chan Soon-Shiong Institute of Molecular Medicine at Windber, PA, USA

***Corresponding Author:** Praveen-Kumar Raj-Kumar, Bioinformatics Scientist at Chan Soon-Shiong Institute of Molecular Medicine at Windber, PA, USA.

Received: August 05, 2017; **Published:** August 21, 2017

Over the past two decades, hybridization-based array methods such as microarray have dominated the field of gene expression profiling [1-3]. The past decade, however, has seen an enormous rise in next generation high-throughput sequencing technologies, particularly whole RNA shotgun sequencing or RNA-Seq [4–6]. Many steps are required in the analysis of RNA-Seq data in order to generate a gene expression profile: from preprocessing such as to remove sequencing contaminants, duplications, low quality bases, poly-A/T tails etc., alignment of sequencing reads to a reference genome or transcriptome, quantification of gene expression and data normalization. This editorial will focus on the common questions encountered when handling a gene expression matrix. Readers are encouraged to go here [7–10] for a comprehensive review/comparison of the various preprocessing methods, alignment tools and data normalization techniques.

Obtaining a true representation of RNA-Seq based gene expression data is not trivial. A simple question such as, "Do I need to log transform the data?" can be tricky. The answer is yes! You need to log transform the data in order to bring it to same scale as it is in microarray [11,12]. Why? Like I stated before, microarray has dominated most gene expression profiling efforts. Hence many bioinformatics software packages like in Bioconductor [13] have been developed to deal with microarray data. The ability of RNA-Seq to give the absolute expression introduces the problem of extreme values and it is heteroscedastic [14,15], meaning variance depends on the mean. Log transformation approximates the data to be homoscedastic which is the assumption for many statistical methods designed for exploratory analysis. Furthermore, to avoid the problem of infinite values, one could choose to use the shifted logarithm (log(x+1)) or the variance stabilized transformation proposed in DESeq [14]. In addition, untransformed data will be dominated by highly expressed genes. For example, untransformed data in principal component analyses (PCA), a commonly-used method to detect technical or sequencing batch effects among the data, will be influenced by highly expressed genes.

Another common question is, "Do I need to standardize (z scores) the data before performing differential expression between the treatments?" The answer is no. The purpose of standardizing data is to bring it to the same scale across the matrix in order to do a clustering analysis that relies upon the distance between genes [14,16]. Moreover, it is recommended to use the Bioconductor packages which are specifically developed to find differential expression among 'count data' like DESeq [14], DESeq2 [15], edgeR [17] and several others [18] provided you have at least three biological replicates per treatment [7] an important point to consider before performing differential expression is filtering of lowly expressed genes. Such filtering avoids the problem of having to consider too many insignificant tests during multiple test correction [19]. Filtering can be done using an approach like that developed in genefilter [19], that is removing genes that have no possibility of detection or any of the genes that have mean expression below an arbitrary threshold.

A common endeavor pursued in this genomics era is to combine publicly available gene expression datasets of the same biological treatments to enhance the statistical power of differential expression analysis. A good approach would be to normalize the data according to its sequencing protocol and then combine the log transformed data [7,11]. After that, a data visualization technique such as a PCA plot, can be used to check for known or unknown batch effects [13,14]. Any analysis of combined gene expression data can only

be performed after adjusting for any known or unknown batch effects [22] using tools like SVA [20], COMBAT [20,21] and ARSyN [23]. Following this adjustment, the data cannot be further used in tools developed for RNA-Seq "count" data as this is no longer count data. Hence appropriate statistical methods implemented in limma [24], SAMr [25], B statistic [26] and many others [27] can be used for the detection of differential expression.

To summarize gene expression counts have to be normalized, log transformed and adjusted for any known or unknown batch effects before performing any downstream analysis like differential expression analysis.

Acknowledgements

I thank Anupama, Lori Sturtz and Hai Hu for critically reading the article.

Bibliography

1. Schena M., *et al*. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". *Science* 270.5235 (1995): 467-470.

2. Brown P and Botstein D. "Exploring the new world of the genome with DNA microarrays". *Nature Genetics* 21.1 (1999): 33-37.

3. Lander ES. "Array of hope". *Nature Genetics* 21.1 (1999): 3-4.

4. Morin R., *et al*. "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing". *BioTechniques* 45.1 (2008): 81-94.

5. Wang Z., *et al*. "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews Genetics* 10.1 (2009): 57-63.

6. Chu Y and Corey DR. "RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation". *Nucleic Acid Therapeutics* 22.4 (2012): 271-274.

7. Conesa A., *et al*. "A survey of best practices for RNA-seq data analysis". *Genome Biology* 17.1 (2016): 13.

8. Dillies M-A., *et al*. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". *Briefings in Bioinformatics* 14.6 (2013): 671-683.

9. Engström PG., *et al*. "Systematic evaluation of spliced alignment programs for RNA-seq data". *Nature Methods* 10.12 (2013): 1185-1191.

10. Kumar PKR., *et al*. "CADBURE: A generic tool to evaluate the performance of spliced aligners on RNA-Seq data". *Scientific Reports* 5 (2015): 13443.

11. Zwiener I., *et al*. "Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures". *PLOS ONE* 9.1 (2014): e85150.

12. van Houwelingen HC., *et al*. "Cross-validated Cox regression on microarray gene expression data". *Statistics in Medicine* 25.18 (2006): 3201–3216.

13. Gentleman RC., *et al*. "Bioconductor: open software development for computational biology and bioinformatics". *Genome Biology* 5.10 (2004): R80.

14. Anders S and Huber W. "Differential expression analysis for sequence count data". *Genome Biology* 11.10 (2010): R106.

15. Love MI., *et al*. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology* 15.12 (2014): 550.

16. Love MI., *et al*. "RNA-Seq workflow: gene-level exploratory analysis and differential expression". *F1000Research* 4 (2015): 1070.

17. Robinson MD., *et al*. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26.1 (2010): 139-140.

18. Rapaport F., *et al*. "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data". *Genome Biology* 14.9 (2013): R95.

19. Bourgon R., *et al*. "Independent filtering increases detection power for high-throughput experiments". *Proceedings of the National Academy of Sciences of the United States of America* 107.21 (2010): 9546-9551.

20. Leek JT., *et al*. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". *Bioinformatics* 28.6 (2012): 882-883.

21. Johnson WE., *et al*. "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* 8.1 (2007): 118-127.

22. 't Hoen PAC., *et al*. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories". *Nature Biotechnology* 31.11 (2013): 1015-1022.

23. Nueda MJ, *et al*. "ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments". *Biostatistics* 13.3 (2012): 553-566.

24. Ritchie ME., *et al*. "Limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* 43.7 (2015): e47.

25. Tusher VG., *et al*. "Significance analysis of microarrays applied to the ionizing radiation response". *Proceedings of the National Academy of Sciences of the United States of America* 98.9 (2001): 5116-5121.

26. Lönnstedt I and Speed T. "Replicated Microarray Data". *Statistica Sinica* 12 (2002): 31-46.

27. Cui X and Churchill GA. "Statistical tests for differential expression in cDNA microarray experiments". *Genome Biology* 4.4 (2003): 210.

Volume 11 Issue 1 August 2017