Journal of
Next Generation Sequencing & Applications

Pareek, Next Generat Sequenc & Applic 2021,
DOI: 10.4172/ 2469-9853.S1-002

Short Communication Open Access

RAP (RNA-Seq Analysis Pipeline)

Chandra S Pareek[1,2]*

[1]*Functional Genomics Lab. Faculty of Biology and Environmental Protection, Nicolaus Copernicus University, Torun, Poland*
[2]*Interdisciplinary Centre of Modern Technology, Nicolaus Copernicus University, Torun, Poland*

Global researches in the field of biology, biotechnology and medicine require utilization of advanced transcriptome analysis to investigate cellular state, physiology and other relevant biological events. In past decade, RNA-seq is emerged as one of the most versatile application of Next-generation Genome Sequencing (NGS) technology and revolutionized the global researches on transcriptome [1]. The high-throughput RNA-seq data continues to provide unparalleled insight into transcriptome complexity [2]. Now one can consider that the "gold standard" for assessing global transcriptome analysis on RNA-seq data is poised to revolutionize our understanding of transcription and posttranscriptional regulation of RNA [3].

Since its first introduction in 2005, tremendous progress has been made both in advancement of NGS platforms (Illumina, Roche, life technologies etc.), as well as in advancement of bioinformatics tools for RNA-seq data analysis. Among the sequencing platforms, Illumina led the way by offering the most versatile equipment, *viz.*, MiSeq, HiSeq series and NextSeq 500/550 [4] for comprehensive transcriptome analysis in mammals. Principally for most of the NGS platforms, the laboratory procedure to generate RNA-seq data is the same involving three major steps: i) isolation of total RNA and/or mRNA ii) library preparations iii) NGS sequencing and production of high-throughput raw sequencing data of transcriptome [5].

However, success of any RNA-seq experiment is critically and solely depends upon the effectiveness of utilized transcriptome analysis methods [6]. That is why with the advancement of novel technologies, the comprehensive nature of generated RNA-seq data has been a boon in terms of transcript identification but the transcriptome analysis is still a big challenge. In general, transcriptome analysis on RNA-seq data were enabled to perform following key tasks: i) detection of rare and novel transcripts (SNPs and *Indels*) [7], ii) determination of the structure of genes (gene counts), iii) quantification of the changing expression levels of individual transcript and its comparison (DEG) [8], iv) identification of splicing variants and other post transcriptional modifications [3], and v) transcriptome mapping or transcript assembly, (de novo assembling and re-sequencing) [9] and vi) transcriptome annotation [10].

To date, an updated review of literature records PubMed Central (PMC) on transcriptome analysis using RNA-seq technology revealed a total 17221 publications. Majority of publications were on human transcriptome analysis (14760). By applying the major applications of RNA-seq as the keywords, publications can be further categorized as: mapping with 9641, assembly with 7651, *de novo* assembly with 4981, splicing variants with 5353, Differentially Expressed Gene (DEG) with 370, rare and novel variant analysis (SNPs and *Indels*) with 3161, and transcript annotation with 8258 publications, respectively.

In general, transcriptome analysis on RNA-seq data starts with raw sequencing data generated from NGS platforms. A single run of any sequencing platform yields an appreciable amount (hundreds of gigabytes) of sequencing data. Before starting into transcriptome analysis on RNA-seq data, one should make sure that the necessary computing, data storage resources and basic bioinformatics expertise are already in place and well set. Generally, RNA-seq raw data come with errors and should be preprocessed before being utilized into downstream analyses such as mapping, assembling etc. At the beginning, the basic tasks such as adapter removal, duplicate quantification and summary statistics on quality score can be performed by standard tools like the fastQC toolkit [11]

Another crucial component for transcriptome analysis of RNA-seq data is the normalization [12,13]. When comparing transcripts of different length, it is important to control for length, as a longer transcript will be covered with more reads than an equally expressed shorter transcript. Therefore, a simple way to handle this issue is to divide read counts by the total number of mapped reads (or quantiles of mapped reads). This basic normalization controlling for transcript length and sequencing effort is captured by the commonly used measures such as RPKM (reads per kilo-base exon per million reads) or FPKM (fragments per kilo-base exon per million reads) [14]. Apart from normalization, it is also important to find a statistical distribution approximating the nature of the RNA-seq data. For this, several software packages such as DE-seq [15], bay-seq [16], edgeR [17], NOI-seq [18], are widely utilized.

A successful transcriptome analysis on RNA-seq data can also yield a set of candidate genes (CGs), SNPs and ESTs databases that differ between treatments or populations. This type of transcriptome analysis particularly has great significance in experiments involving active breeding populations of plants and animals. For instance, most of the farm animals and plants, CGs, SNPs and ESTs of economic traits and the trait-associated QTLs (Animal qtldb) [19], SNPs and ESTs database are well set and updated.

Apart from its great potential and versatility, transcriptome analysis on RNA-seq data still has some drawbacks for instances, i) the existence of artifacts and biases (regional GC content, preferential sites of fragmentation, and read "pile-up" due to primer affinity and transcript end effects) need to be identified and controlled for, ii) although many methods have been developed to effectively analyze the RNA-Seq data, but improvements are needed to deal with problems associated with multi-reads and estimating the abundance of splice variants.

In a recent developments on transcriptome analysis on RNA-seq, research communities are designing, developing and offering the users' friendly and open-access cloud computing web resources [20] that can cover tool installation, relevant file formats, reference genomes, mapping, assembly, transcriptome annotations, quality-control strategies, expression, differential expression, and alternative splicing analysis methods. Another multi-task cloud computing can also be performed by implementing a complete but modular analysis workflow pipeline termed as RAP (RNA-Seq Analysis Pipeline). Using

***Corresponding author:** Chandra S. Pareek, Functional Genomics Lab. Faculty of Biology and Environmental Protection, Nicolaus Copernicus University, Torun, Poland, Tel: +48 56 611-26; E-mail: pareekcs@umk.pl

Received December 10, 2021; **Accepted** December 28, 2021; **Published** January 31, 2021.

Citation: Pareek CS (2021) RAP (RNA-Seq Analysis Pipeline) 7:1. doi:10.4172/2469-9853.S1-002

RAP cloud computing, the user can perform a complete transcriptome analysis without any specific technical competence as well as by directly managing the complexity of distributed computational resources. Moreover RAP cloud computing offers a web interface for results management and visualization, by allowing the user to browse and filter the massive amount of data obtained from typical RNA-Seq experiments [21]. With the availability of more and more increasing volume of transcriptome analysis data from large-scale RNA-Seq studies [22], developed ″Stormbow″ a cloud-based software package, to process large volumes of RNA-Seq data in parallel. Similarly, the user-Friendly gene eXpression analytic tool (FX), which also runs in parallel on cloud computing infrastructure, was developed for the cloud computing of gene expression levels and genomic variant calling [23]. Finally, one can conclude that global transcriptome analysis on RNA-seq data continuously evolve and within the next few years, it will without doubt be exploited to a larger extent and lead to many more innovating discoveries to understand the complexity of mammalian transcriptome.

References

1. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet. 52: 413-435.
2. Calabrese C, Mangiulli M, Manzari C, Paluscio AM, Caratozzolo MF, et al. (2013) A platform independent RNA-Seq protocol for the detection of transcriptome complexity. BMC Genomics 14: 855.
3. Spangenberg L, Shigunov P, Abud AP, Cofré AR, Stimamiglio MA, et al. (2013) Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes. Stem Cell Res 11: 902-912.
4. http://www.illumina.com/systems/sequencing-platform-comparison.html
5. Pareek CS (2014) An overview of next generation genome sequencing platforms. In Next Generation Sequencing: Current Technologies and Applications. Caister Academic Press, Canada.
6. Hoeijmakers WA, Bártfai R, Stunnenberg HG (2013) Transcriptome analysis using RNA-Seq. Methods Mol Biol 923: 221-239.
7. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, et al. (2013) Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One. 8:e58815.
8. Rajkumar AP, Qvist P, Lazarus R, Lescai F, Ju J, et al. (2015) Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics 16: 548.
9. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12: 671-682.
10. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.
11. Kroll KW, Mokaram NE, Pelletier AR, Frankhouser DE, Westphal MS, et al. (2014) Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline. Cancer Inform 13: 7-14.
12. Zhou Y, Lin N, Zhang B (2014) An iteration normalization and test method for differential expression analysis of RNA-seq data. BioData Min 7: 15.
13. Filloux C, Cédric M, Romain P, Lionel F, Christophe K, et al. (2014) An integrative method to normalize RNA-Seq data. BMC Bioinformatics 15:188.
14. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR (2013) Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14: 778.
15. Tang M, Sun J, Shimizu K, Kadota K (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC Bioinformatics 16: 361.
16. Hardcastle TJ (2012) Empirical Bayesian analysis of patterns of differential expression in count data. R package version 2.4.1.
17. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.
18. Khang TF, Lau CY (2015) Getting the most out of RNA-seq data analysis. PeerJ 3: e1360.
19. http://www.animalgenome.org/cgi-bin/QTLdb/index
20. Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLoS Comput Biol 11: e1004393.
21. D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, et al. (2015) RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. BMC Genomics 16: S3.
22. Zhao S, Prenger K, Smith L (2013) Stormbow: A Cloud-Based Tool for Reads Mapping and Expression Quantification in Large-Scale RNA-Seq Studies. ISRN Bioinform 2013: 481545.
23. Hong D, Rhie A, Park SS, Lee J, Ju YS, et al. (2012) FX: an RNA-Seq analysis tool on the cloud. Bioinformatics. 28: 721-723.

This article was originally published in a Volume 7:issue1
handled by Editor(s). Dr. Jianping Wang, University of Florida, USA.