Xournals

Academic Journal of Bioinformatics

ISSN UA | Volume 01 | Issue 01 | January-2019

A Study on Differential Gene Expression by RNA-Sequencing Technology

Mahesh Sharma[1] and Pradyut Kumar Mohanty[2]

Available online at: www.xournals.com

Received 18th September 2018 | Revised 30th October 2018 | Accepted 19th December 2018

Abstract:

Differential gene expression is an important project in the field of biology. From the study of gene expression, the production of new gene product can be made by the information of gene. These expression need very high skill and time. In past, gene expression is done by the microarray based techniques that have the many drawbacks. Nowadays new technology 'RNA sequencing (RNA-Seq) is used for the gene expression that overcomes the drawback of microarray techniques. In RNA-Seq analysis, a large number of tools are accessible that have the same steps: reading the alignment, expression modeling, and determination of variably expressed genes. These tools are edgeR, DESeq, baySeq, NOIseq and so on. This paper represents the presentation of RNA-Seq in the identification of differential gene expression with their different software tool. It also defines the types of RNA-Seq that have the use in the detection of gene of disease i.e., Cancer.

***Keywords:** Gene Expression, RNA-Seq, Tools*

Authors:

1. *Web Univ, Sikkim Manipal University Distance Learning Centre, South Extension, Delhi, INDIA*
2. *Prof. Web Univ, Sikkim Manipal University Distance Learning Centre, South Extension, Delhi, INDIA*

Introduction

For the analysis of gene expression, RNA sequencing (RNA-Seq) has a number of technological benefits such as, a wider dynamic range and the liberty from predesigned probes. This RNA sequencing examine transcriptomes and can be applied in biological research, drug discovery and clinical development. RNA sequencing avoids some of the technical limitations such as varying probe performance and cross-hybridization and have broader dynamic range in the comparison of microarray-based transcriptome profiling. In biological system, the expression levels of thousands of genes can be measured simultaneously by the help of RNA-Seq that also provides insights into functional pathways, regulatory networks, alternative splicing, unannotated exons and novel transcripts. In the analysis of gene expression, the gene expression signatures changes are identified by the comparison of two or more condition **(Williams et al, 2017)**.

Types of RNA Sequencing

Single RNA-Seq: In single cell, RNA sequencing have a new approach with the study of complex biological processes. In recent years, qualitative microscopic images and quantitative genomic datasets are used by single RNA-Seq for the study of cancer. There are many disease can be resolved such as resolving solid tumor heterogeneity, recognizing stem cells, tracing cell lineages and population consumption, measuring the mutation rates, and detecting the fusion gene sequences by the single cell genome and exome sequencing. In this way, single cell sequencing provides more accurate measurement.

Dual RNA-Seq: The response of eukaryotic cells is another significant field where the RNA-Seq plays a vital role. In the examination of Transcriptoms, the main focus is either on the host or the pathogen that needs the RNA molecule segregation from the host or pathogen at a specific point of time. By the help of dual RNA-Seq, the gene is monitored from the host or pathogen without RNA segregation throughout the infection process. In many such areas as molecular and cellular biology, immune response, public health in disease, bacteria and plant interaction, Dual RNA-Seq technique is used.

Differential Gene Expression

For the differential expression analysis, RNA-Seq is very helpful that involve some specific conditions with five steps.

1. Small complementary DNA (cDNA) sequences are formed form the RNA samples and sequenced form a high throughput stage.
2. Small complementary DNA structures are plotted to a genome or a transcriptome.
3. For each single gene or isoform, the expression levels are predicted.
4. Then, plotted data are standardized and differential expressed genes (DEGs) are identified by arithmetical and machine learning methods.
5. In last, the formed data relevancy is estimated from biological setting.

Various software and pipelines were developed for the examination of gene expression from RNA Sequencing technology.

For a general gene expression analysis, RNA sequencing is categorized into two chief subcategories: parametric and non-parametric. In parametric systems, all information related data is captured inside the parameters. From these mentioned parametric methods, the charge of this unknown data can be predicted from the observation of implemented design and its parameters. During the use of parametric methods in the analysis of gene expression, it is assumed that each appearance value for a gene is plotted into a specific dispersal after a normalization i.e., poisson, and also called negative binomial. The negative binomial distribution or poisson distribution is also termed as the gamma-Poisson distribution that is an overview of the Poisson distribution which allow for an additional modification.

In case of non-parametric approaches, additional particulars are captured about the data dispersal for example not imposing a firm model to be fixed. This is due to non-parametric replicas taken into considered as the finite set of parameters cannot define the data distribution. Hence, the increment in the quantity of material about data is possible within its volume.

Some tools like as edgeR and baySeq use the undesirable binomial model in the RNA-Seq variance manifestation analysis. While NOIseq and SAMseq tools use non-parametric methods. Other these tools, some of the methods are built on transcript discovery that were established to recognize indefinite transcripts or isoforms. From the transcript detection, the documentation of DEGs such as EBSeq and Cuffdiff2 is possible (**Silva, Domingues and Lopes, 2017**).

edgeR: It is used to determine the differential expression by the help of experimental Bayes assessment and precise tests that completely rely on negative binomial model. The entire moderation of degree of over-dispersion across genes is done by the Bayes procedure that borrow information between genes.

DESeq: It is generally based on the negative binomial distribution similar to the edgeR. This tool observe the rapport between the mean and variance during the estimation of dispersion. By this method, a well composed selection of variably expressed genes is possible throughout the vibrant series of data. It also allows the analysis of tests with small numbers of replicates and can be work without any biological replicates.

baySeq: It uses the Bayesian empirical approach for the estimation of posteriori probability of each set of models. This tool assumes negative binomially distributed data. For library scaling factors, various library sizes are taken into account. This method is computationally intensive but compare to others, the advantage of parallel processing can be taken by its implementation.

NOIseq: It is also a non-parametric method in which contrasting fold-change differences and absolute expression differences among the sample help in the empirical models for the noise distribution from the actual data. According to developer of method, the rate of false discoveries can be controlled and the size of data set can be adapted.

SAMseq: It is a non-parametric method that use re-sampling procedure for the sequencing counts with different depths. According to developer of method, this tool can be applied to data that have at least moderate numbers of replicate samples. This method enables to select significant features effectively compare to the parametric methods in case of unmanaged distributional assumptions.

Limma: It use the linear model that was developed to analyze data from microarray but in recent RNA-Seq analysis is used. The use of TMM normalization of the edgeR package called ‘voom’. In this tool, Benjamini-Hochberg procedure is used to estimate the FDR.

Cuffdiff 2: At transcript level resolution, the gene expression is estimated and it controls variability and read mapping ambiguity by the help of beta negative binomial model for the fragment counts. Cuffdiff2 is the part of extensive Cufflinks package that is developed to identify the differentially expressed genes. Cuffdiff2 analyze the signals at transcript level and gives report on differential expression at gene level. For the comparison with other software packages, these gene level results are used.

EBSeq: In this tool, empirical Bayes autoregressive hidden Markov model is applied for the identification of dynamic gene in two steps. In first step, Negative binomial (NB) model is used to estimate the parameters. Then in second step, Gene is categorized at each time point by the help of Markov-switching autoregressive model. After that gene is classified into expression path **(Silva, Domingues and Lopes, 2017; Seyednasrollah, Laiho and Elo, 2013, Spies et al, 2017)**.

Review of Literature

Silva, Domingues and Lopes (2017) discussed the documentation of variably expressed gene or transcripts in which they appraised the effect of six mapping methods and nine methods for the DEGs identification. In their paper, DESeq2, NOIseq and limma methods showed the distinct result 95%, 95% and 93% respectively. They recognized that with the amalgamation of five techniques, the establishment is done with the high sensitivity and give more reliable results.

William *et al* (2017) in their paper, they take the heterogeneous samples for the identification of gene expression. After the analysis, they found that the choice of RNA-Seq technology was a significant. They also determined that the influence of software selection at each specified step was not a purpose of upstream position. They gave a suggestion that the choice of workflow should be depend on how will be result used.

Huang, Niu and Qin (2015) proposed the method for identification of a gene expression. They discussed about RNA-Sequencing is a new and rapidly growing technique in the field of research. The negative binomial distribution in parametric framework is most common assumption rather than the poisson distribution because of the technical and biological variations. There are different types of tools for the estimation of gene expression even in the case of small number of replicates.

Han *et al* (2015) in their review paper, they showed the application of RNA-Sequencing in a biomedical examination highlights the in-depth computational methodology in data preprocessing, variable gene expression, alternative splicing, path analysis, and co-expression linkage by which the understanding of genomic level can be increased.

Finotello and Camillo (2014) stated that the study of variance gene expression can be determined by the RNA-Sequencing technology. In the RNA-Sequencing data, the description of a computational pipeline is required includes the numerous steps; recite mapping, sum total the computation, normalization and testing for variance gene expression. According to researcher, RNA sequencing is developing at very high rate and used as a third generation technologies.

Spies *et al* (2017) discussed the tools of RNA-Seq and their combination for the estimation of gene expression. They concludes that time points is a robust and accurate approach on experiment setup. In the end of their paper, they said that several methods combination is the most reliable and cost-effective for the replicates increment or time point.

Conclusion

RNA sequencing technology is an advanced technique that is best for the differential gene expression. This technology use many different types of software for the analysis. These software can be used individually as well as in the combination and give the accurate and reliable result. This paper concludes that the use of RNA sequencing technology with the combination can provide a better result compare to individual and also can be used to determine the diseases.

References:

Conesa, Ana, *et al.* “A Survey of Best Practices for RNA-Seq Data Analysis.” *Bio Medical Science Central*, 2016, pp. 1–19.

Finotello, F., and B. Di Camillo. “Measuring Differential Gene Expression with RNA-Seq: Challenges and Strategies for Data Analysis.” *Briefings in Functional Genomics*, vol. 14, no. 2, 2014, pp. 130–142.

Han, Yixing, *et al.* “Advanced Applications of RNA Sequencing and Challenges.” *Bioinformatics and Biology Insights*, vol. 9s1, 2015.

Huang, Huei-Chung, *et al.* “Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software.” *Cancer Informatics*, vol. 14s1, 2015.

Silva, Juliana Costa, *et al.* “RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool.” *PLOS ONE*, vol. 12, no. 12, 21 Dec. 2017, pp. 1–18.

Spies, Daniel, *et al.* “Comparative Analysis of Differential Gene Expression Tools for RNA Sequencing Time Course Data.” *Briefings in Bioinformatics*, June 2017.

Williams, Claire R., *et al.* “Empirical Assessment of Analysis Workflows for Differential Expression Analysis of Human Samples Using RNA-Seq.” *BMC Bioinformatics*, 2017, pp. 1–12.