Macrothink Institute™

Journal of Biology and Life Science
ISSN 2157-6076
2017, Vol. 8, No. 2

An Overview of RNA-seq Data Analysis

Tahsin Ferdous

Department of Statistics,

Shahjalal University of Science and Technology,

Bangladesh. E-mail: tahsinferdous29@gmail.com

Mohammad Ohid Ullah (Corresponding author)

Department of Statistics,

Shahjalal University of Science and Technology,

Bangladesh. E-mail: ohid-sta@sust.edu

Received: May 22, 2017 Accepted: June 8, 2017 Published: August 3, 2017

doi:10.5296/jbls.v8i2.11255 URL: https://doi.org/10.5296/jbls.v8i2.11255

Abstract

Latest breakthrough in high-throughput DNA sequencing have been launched different arenas for transcriptome analyses, jointly named RNA-seq (RNA-sequencing). It exposes the existence and amount of RNA in a biotic sample at a specific time by utilizing next generation sequencing (NGS). In this review, we aimed to explore the several methods which are applied in analyzing RNA-seq data. We also discussed its importance over microarray data. As establishment of several methods have already taken place to analyze RNA-seq data, therefore, further analysis is very essential to select the best one to avoid false positive outcomes.

Keywords: RNA-seq, transcriptome, Methods, Models, NGS

1. Introduction

1.1 Gene Expression

The information contained within a gene turns into an effective product by gene expression. Genes can be expressed as RNA and translated into protein; expression arises one at the transcription level, in which RNA is produced from DNA, and one at the protein level, where protein is created from mRNA Several different steps are included through which DNA is transcribed into RNA and this in turn is modified into a protein or in some cases an RNA

(Cheriyedath, 2016).

Genes control the production of proteins in a biological system through transcription and translation. Gene regulation is essential as the rate and manner of gene expression is controlled by it.

Gene expression can be measured by several techniques including serial analysis of gene expression (SAGE), Complementary DNA (cDNA) subtraction, differential display, RNA sequencing and microarrays. Of them, microarrays and RNA sequencing have extensive application in gene expression analysis.

1.2 Microarrays

The genomics background for the word “microarray” usually specifies a device where single-stranded DNA oligonucleotides or “oligos” are attached to a compact exterior. As a result of its tendency of being double stranded, a sample staying in accurate buffer, is attached to the exterior part of the microarray. The inactive complementary DNA oligo will be combined with the independent swollen samples. Relying on this property, a fluorescent dye is either attached before to sample inclusion and hybridization or after the DNA hybridization to the microarray. One or two fluorescent dyes can be utilized before to sample inclusion. For conducting gene expression analysis of thousands of genes in this circumstance, a microarray is a floor of high-throughput DNA or RNA hybridization. Also, the whole genome can be covered by it (Wang et al.,2012).

1.3 Gene Expression by Sequencing

The progressive DNA sequencing technology is another extensively used method for exploring transcriptomes. Since the middle of 1990s, microarray was preferred for analyzing gene expression, but the stage for sequencing to be a captivating substitute technology for biological research is rapidly set by the Sanger sequencing biochemistry (Sanger et al., 1977).

Fred Sanger developed the DNA sequencing method which now constructs the basis of automated "cycle" sequencing reactions. For two major advancements in the 1980s, researchers thought that the whole genome sequencing could be possible. A skillful procedure called polymerase chain reaction (PCR) was the first that allowed many copies of DNA sequence to be rapidly and properly produced. A converted automatic method of DNA sequencing which was built upon the chemistry of PCR was the second. Frederick Sanger developed the sequencing process in 1977.

In 1990, the most vital step in the genome sequencing of higher organisms launched with the Human Genome Project (HGP) with the aim of complete mapping and realizing of all of the genes of human beings (Lander et al., 2001; Venter et al., 2001). As a consequence of massive success of HGP, the analysis of many complications regarding biology, disease and the environment is now possible by a very large-scale sequencing of genome. Recently, various next generation sequencing platforms are accessible including 454-FLX (Roche) (McGill University and Génome Québec Innovation Centre, 2014), the Genome Analyzer (Illumina/Solexa) (Illumina, 2009), and SOLiD (Applied Biosystems) (Life Technologies

Corporation, 2010). All these NGS platforms are based on parallelizing the sequencing process (Hutchsion, 2007; Pettersson et al., 2009; Shendure and ji, 2008). RNA-seq is the most familiar application for next generation sequencing.

1.4 RNA Sequencing

RNA-seq (RNA-sequencing) exposes the existence and amount of RNA in a biotic sample at a particular moment by utilizing next generation sequencing (NGS) (Chu and Corey, 2012).

A RNA sequence experiment creates a group of cDNA fragments in all cases. At first a sample of purified RNA is cut and transformed into cDNA. Utilizing the short-read sequencing this collection of cDNA is then sequenced on a high-throughput platform such as Illumina, SOLiD or Roche454. A large number of short sequence reads that coincide to distinct cDNA fragments are generated by this short-read sequencing (Oshlack et al., 2010). A normal RNA-seq to be comprised of the steps named as design experiment, RNA preparation, library preparation, sequencing the cDNAs and analyzing the resulting short read sequences.

An outline of RNA-seq experiment can be understandable by the following figure:

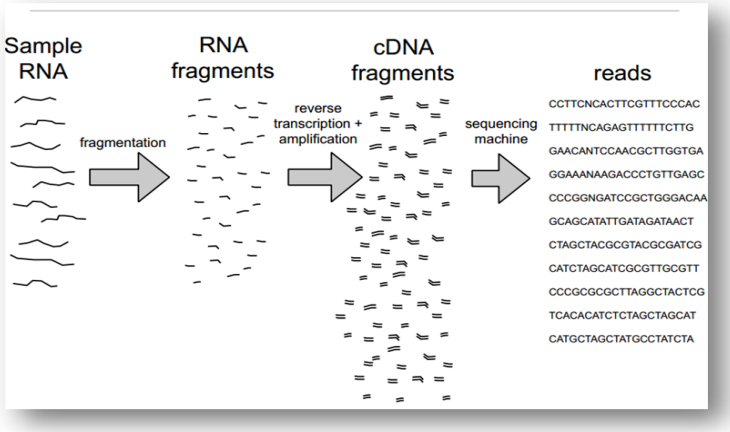

Figure1. An outline of RNA-seq experiment

The most common use of RNA sequencing is in the search for differentially expressed (DE) genes, that is, genes that exhibit differences in expression level between several conditions. In RNA-seq data analysis procedure, at first the images from sequencing reactions are captured and bases are arranged in FASTQ file format. These reads are then particularly mapped to a reference genome or transcriptome with the help of different alignment tools. Next the data are normalized by normalization process and then calculation of gene expression is conducted. Researchers can further proceed to gene discovery, transcript abundance and also alternative splicing.

The type of differential gene expression is estimated by using statistical distributions. Poisson and Negative Binomial (NB) distributions are the two most extensively used distributions for fitting RNA-seq data. Some other distributions such as the Beta-binomial, have also been

suggested (Anders and Huber, 2010).

1.5 Microarrays vs. RNA-seq

In gene expression studies, microarrays and sequence-based methods are frequently used together, with a rising fame of the application of RNA-seq over microarrays in transcriptome analyses. The both platforms exhibit high one to one reproducibility and there is high correlation ranging from 0.62-0.75 between gene expression profiles created by the two methods (Fu et al., 2009). Intensities are measured by using continuous distribution in array-based technology, while RNA-seq provides discrete measurement of reads for each gene.

Microarray technology confines the researcher to finding transcripts that correspond to existing genomic sequencing information. On the other hand, RNA-seq is perfect for discovery-based experiments by working efficiently for inspecting both known transcripts and searching new ones. For improved detection of genes, transcripts, and differential expression, RNA-seq possesses higher sensitivity and dynamic range of expression levels, with absolute rather than relative values, with lower technical variation and thus higher precision than microarrays. That's why transcriptome studies are shifting to depend on sequencing-based methods rather than microarrays (Robinson and Smyth, 2007). Measuring expression levels in digital, in place of analog is one more benefit of RNA-seq (Matukumalli and Schroeder, 2009).

In microarray analysis, it becomes tough to make lay-out of certain probes pointed at particular sequences because of the binding affinity constraint which makes part of the genome to be unreachable (Bradford et al.,2010). DNA sequences can be exactly mapped to particular areas of the genome. As a result, low background signal is provided by RNA-seq and noise in the experiment is simply omitted during analysis. Hybridization related matters which are observed in microarrays, are also removed in RNA-seq experiments. Thus, another signal-to-noise prevalence is presented (LaFranzo, 2013). Quantification of individual transcript isoforms is another advantage of RNA-seq (Malone and Oliver, 2011).

RNA sequencing analysis is the most modern technology although it has some limitations and biases. Data storage and analysis of this sequencing process is more harder with no standard protocol which needs more time than any microarray technology. Despite of these, RNA-seq analysis is turning into an effective transcriptome profiling tool day by day as it has made possible the qualitative and quantitative advancements to gene expression analyses in a cost-effective way.

1.6 Software Packages for Detecting Differential Expression

For the detection of differential expression, some software packages are generally used in RNA-seq analysis. The most widely used software packages are summarized by the following Table 1. The default normalization method is underlined in the Table, when various normalization methods are available.

Table 1. Most widely used software packages which are applied in determining differential expression in RNA-seq.

	Normalization	Model	Differential expression test	Ref.
DEGseq	'None', 'loess' and 'median'.	Poisson model	Z score test, Fisher's exact test (FET), large sample approximation, such as the LRT.	(Wang et al.,2010)
edgeR	TMM/Upper quartile/RLE (DESeq-like)/None	Negative-binomial model	Exact test	(Robinson et al.,2010)
DESeq	DESeqsizeFactors	Negative-binomial model	Exact test	(Anders and Huber,2010
baySeq	Scaling factors (quantile/TMM/total)	Negative-binomial model	On the basis of the natural logarithmic scale, the estimated posterior likelihoods are notified.	(Hardcastle and Kelly,2010)

2. Methodologies

2.1 Correctional Techniques

Filtering and normalizing the data are the two correctional methods which are used in RNA-seq analysis.

2.1.1 Filtering

In RNA-seq analysis, we use filtering criteria to reduce the consideration of differentially expressed genes whose expression levels could be tolerably hypothesized to be below the level essential to affect cellular function or phenotype. Thus, we are permitted to emphasize on those that are most acceptable to be biologically significant (Manthey et al., 2014).

2.1.2 Normalization

Normalization is an imperative step in the RNA-seq data analysis. The amount of reads which are monitored because of a gene relies on the expression level and the length of the gene, and also on the RNA composition of the sample (Balwierz et al., 2009; Oshlack and

Wakefield, 2009). The effect of gene length and total sample RNA composition is lessened by the normalization process. As a result, a direct clear representation of the targeted gene expression level is shown by the normalized read counts. Several normalization methods are used in RNA-seq analysis. Some of them are Total Count (TC), Upper Quartile (UQ), Median (Med), DESeq, Quantile (Q) and Reads PerKilobase per Million mapped reads (RPKM) (Dillies et al.,2013).

2.2 Different Methodology Used in this Analysis

Several tests along with several software packages are applied in the recognition of differentially expressed genes in RNA-seq analysis. We will discuss some of these methods which are most commonly used.

2.2.1 Fold Change

Fold change is the simplest method to classify genes as differentially expressed. In this method log ratio between two conditions or the average of ratios when there are replicates among the two conditions are calculated and a particular gene is said to be differentially expressed for which the fold change value varies by more than a random cut-off value. However, this test is not a statistical test, it is preferred by biologist for its simplicity and it gives a sense about the difference between two conditions. The fold-change for gene i is defined as,

$$FC_i = \bar{x}_i - \bar{y}_i \text{ or, } FC_i = \frac{\bar{x}_i}{\bar{y}_i}$$

Where, $\bar{x}_i$ and $\bar{y}_i$ are the means of the two groups' raw expression values x_{ij} and y_{ij}, respectively. x_{ij} is the raw expression levels of gene i in replicate j in the control group and y_{ij} is the raw expression levels of gene i in replicate j in the treatment group.

The procedure is to compute the log ratio between the expression levels in two conditions and identify genes as differentially expressed whose ratio exceed an arbitrary cut-off value(for instance, 2- fold). In fold change method, genes with large variances are vulnerable to make the cutoff easily just because of noise. So it is probable to have genes with large fold change which actually are not statistically significant just because the populations show much variability. Similarly, it's also possible to have genes with small fold changes which are highly statistically significant because the populations show slight variability.

2.2.2 Robinson and Smyth Exact Test

Robinson and Smyth Exact Test is similar to Fisher's exact test in the case of contingency tables. But here the hypergeometric probabilities are replaced by Negative Binomial. This test is only appropriate for the experiments with a single factor (Robinson and Smyth, 2008).

2.2.3 Likelihood Ratio Test

Likelihood ratio test is used in many RNA-seq algorithms for the detection of differential expression. The likelihood of the data considering no differential expression (null model) against the likelihood of the data considering differential expression (alternative model) is

compared by this test.

$$D = -2\log \frac{likelihood\ of\ null\ model}{likelihood\ of\ alternative\ model}$$

Where, D follows a $\chi 2$ distribution which can be applied to compute a p value.

2.2.4 Controlling False Discovery Rate (FDR)

RNA-seq data analysis involves in testing tens of thousands of genes simultaneously for differential expression in one study, hence it requires to testing multiple hypothesis simultaneously. Analyzing several hypotheses instantaneously increases the rate of type I error. To handle this problem several multiple testing correction methods have been developed, which are relied on family-wise error rate (Noble,2009) and False Discovery Rate (FDR) (Benjamni and Hochberg,1995).

For the control of FDR several methods have been suggested including, Benjamini and Hochberg FDR, Storey's positive FDR (pFDR) (Storey, 2003; Storey, 2003). Among these methods, the Benjamini and Hochberg FDR is the most flexible method compared to other methods and it is widely used in RNA-seq data analysis. Also, it is easily available with all statistical packages.

2.3 Model

Poisson and negative binomial (NB) are two more sensible models of differential expression in RNA-seq data.

2.3.1 Poisson Model

Primitive RNA-seq studies using only technical replicates informed that Poisson distribution is suitable to the counts for most of the genes. The main benefit of this distribution is its simplicity. According to further studies, Poisson assumption could not be able to identify biological variability, resulting in high false-positive rates because of neglecting the sampling error (Anjum et al., 2016).

2.3.2 Negative Binomial Model

The negative binomial distribution permits modeling of more general mean–variance relationship. Specifically for genes expressed at a higher level, the variance of counts is normally greater than their mean. This criterion is called "over-dispersion". Negative binomial distribution models considering over-dispersion is the most appropriate for the distribution of read counts over biological replicates (Kukurba and Montogomery, 2015).

2.4 Computing Tools

A short brief of most common R software packages which are widely used in finding differentially expressed genes in RNA-seq analysis are given here.

2.4.1 DEGseq

Based on the Poisson model, the R package DEGseq was used to detect differentially expressed genes for RNA-seq data and it is very simple to use. The model considers that the log-ratios of the different biological samples data possess a normal distribution, conditional on the log geometric mean of the data.

'None', 'loess' and 'median' are the different choices for normalization in DEGseq. Among them, ‘none’ is the suggested method.

Fisher's exact test (FET), the likelihood ratios test (LRT) and samWrapper are the three existing methods in DEGseq package. Among them, samWrapper was already established for the analysis of microarray data (Tushar et al., 2000).

Z score test, Fisher's exact test (FET) and large sample approximation, such as the LRT are the preferences. Applying the methods of either Benjamini and Hochberg (Benjamini and Hochberg, 1995) or Storey and Tibshirani (Storey and Tibshirani, 2003), multiple testing was modified.

2.4.2 edgeR

edgeR is a R Bioconductor package which is outlined for the analysis of differential gene expression using replicated count-based expression data under a negative binomial model and established by Robinson and Smyth(Robinson,2010). The package is highly manageable. In this package, the negative binomial model is capable to isolate biological from technical variation. Empirical Bayes methods are applied by this approach to regulate the degree of over-dispersion (Smyth, 2004).

The software package edgeR follows negative binomial distribution. Assume, the observed data is denoted by Y_{gij}. Where, gene (tag, exon, etc.) is denoted by g, experimental group is denoted by i and the index of samples is denoted by j.
We can model the read counts as,
$Y_{gij} \sim NB(M_{jpgi}, \phi_g)$, with mean $\mu_{gi} = M_{jpgi}$ and variance = $\mu_{gi}+\mu_{gi}^2 \phi$. Here, the library size (the sum of the counts of tags in a sample) is denoted by M_j and the proportion of tag g of the sequenced sample for group i is denoted by p_{gi} .Biological or sample-to-sample variation is assessed by the over-dispersion parameter ϕ_g which is related to Poisson. The Negative binomial distribution converts to Poisson distribution if the over-dispertion parameter ϕ_g =0.

For the Negative Binomial distribution, an exact test build upon the normalized data is applied by edgeR which is equivalent to Fisher's exact test (FET) although adjusted for over-dispersed data. Pairwise comparisons of groups are permitted by the ‘exact Test’ function. LogFC that is the log-fold change difference in the counts between the groups, and exact p-values are added by one of the objects which is created by this function.

2.4.3 DESeq

Anders and Huber (Anders and Huber, 2010) develop the Bioconductor package DESeq by adjusting the relationship of mean and variance of the over-dispersed model applied in edgeR. The model used in DESeq is based on the Negative Binomial distribution. Various coefficients of variation for various expression strengths are used in DESeq to estimate the

variance in a local style. As a result, inherent selection biases existing in the hit list of differentially expressed genes are eliminated. Thus, more stabilized and appropriate outcome is achieved.

Now generalized linear models (GLMs) are ready to apply for DESeq. The model is:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\text{With} \quad \log \frac{\mu_{ij}}{s_j} = \eta_{ij} = \beta_{i0} + \sum_l \beta_{il} x_{jl}$$

Where, K_{ij} is read count for gene *i* in the sample *j*, s_j **is size factor for sample *j*,** α_i is dispersion for gene *i*, μ_{ij} is expectation for gene *i* in sample *j*, η_{ij} is linear predictor for gene *i* in sample *j*, x_{jl} is *l*-th predictor for sample *j* (indicator or quantitative), β_{il} is *l*-th regression coefficient for gene *i*. Where, log link, with sample-dependent size factors is the link function and the family is negative binomial with known dispersion. The negative binomial belongs to the exponential family if the dispersion is provided.

We test for differential expression by computing a p-value that exhibits the probability of the null hypothesis. P-values are calculated through a method that is equivalent to a Fisher's exact test, using a 2x2 contingency table. With large samples, a chi-squared test can be used in this situation. More precisely, a χ 2 likelihood ratio test computes the p-values to fit GLMs of the negative binomial family with log link (Anders and Huber, 2016).

2.4.4 baySeq

Hardcastle and Kelly (2010) established an empirical Bayesian analysis approach to determine if there is differential expression between two different conditions (Hardcastle and Kelly, 2010) which gives permission for analyzing data for more complicated experimental designs. It starts by considering that the data follow a distribution, either Poisson or Negative Binomial (NB), which is specified by a set of latent parameters. Two hypotheses are visualized for each gene or tag. One of them assumes no differential expression between the two conditions for the gene, while the other assumes differential expression for the gene. If the prior estimates and the likelihood of the distribution of the data are provided, then one will be able to estimate the posterior likelihood under the two hypotheses to detect if there is differential expression (DE) for that gene. In general, baySeq suggests us to apply the Negative Binomial model. The requisite data format is similar to the edgeR package. Parallel processing is rendered through the 'snow' package for quicker processing.

In baySeq package, no normalization process is suggested.

The package employs two models. One of them is to consider a Poisson distribution on each tag that is $Y_{gij} \sim (M_{jpgi},)$, where the prior for p_{gi} is considered to follow gamma distribution $p_{gi} \sim \Gamma(\alpha_{gi}, \beta_{gi})$. Therefore, the model is titled as the Poisson-gamma approach. The data are Negative Binomial distributed, $Y_{gij} \sim NB\ (M_j p_{gi}, \phi_g)$ which is considered by other models.

On the basis of the natural logarithmic scale, the estimated posterior likelihoods are notified for this package.

3. Conclusions

Among these software packages mentioned above, DEGseq is the simplest to apply. baySeq needs much longer to run with the suggested number of iterations for the rebooting.

Over-dispersed data which is very ordinary among biological samples cannot be tackled by DEGseq. The estimates of the over-dispersion parameter are rendered by the other packages based on Negative Binomial distribution. DESeq and edgeR are the methods which are both based on the negative binomial distribution. If we want to find more significantly differentially expressed genes, then edgeR was much faster than DESeq. Although it took somewhat longer time to estimate the dispersion. DESeq is preferred if the amount of false positives is a main interest. Otherwise, edgeR is lightly better to conduct differential expression analysis instead of probably bringing more false positives (Zhang et al.,2014).

edgeR is the most adjustable package. It can manage both Poisson data and over-dispersed data without the necessity of pre-identifying the model. These two models are also contained in baySeq but one has to pre-identify which to apply. Simulations and real data analysis illustrate that the baySeq works well compared to the other methods in the analysis of pairwise differential expression (Hardcastle and Kelly,2010).

In this review, we tried to discuss some packages to analyze RNA-seq data. However, it's more important to find out the best technique to analyze such kind of data. Therefore, in future more comparative research is essential.

References

Cheriyedath, S. (2016). Gene Expression: An Overview. [Online] Available: http://www.news-medical.net/life-sciences/Gene-Expression-An-Overview.aspx

Wang, J., Tan, A., C., & Tian, T. (2012). Next Generation Microarray Bioinformatics: Methods and Protocols.

Sanger F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74*, 5463-5467.

Lander, E., Linton, L., …, & Doyle, M. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*, 860-921.

Venter, J., Adams, M., ..., & Evans, C. (2001). The sequence of the human genome. *Science, 291*, 1304-1351.

McGill University and Génome Québec Innovation Centre. (2014). User Guide: Roche 454 sequencing technologies.

Illumina, Inc. (2009). Illumina LIMS. User Guide.

Life Technologies Corporation. (2010). Applied Biosystems SOLiDTM 4 System SETS Software. User Guide.

Hutchison, C. A., III. (2007). DNA sequencing: bench to bedside and beyond. Nucleic Acids Research 35:6227-6237.

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technolo-gies. *Genomics, 93*, 105-111.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*(10), 1135-1145.

Chu, Y., & Corey, D. R. (2012). "RNA sequencing: platform selection, experimental design, and data interpretation". *Nucleic Acid Ther, 22*(4), 271–4.

Oshlack, A., Robinson, M. D., & Young, M. (2010): From RNA-seq reads to differential expression results. *Genome Biol2010, 11*, 220.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, R106.

Fu, X., Fu, N., ..., & Khaitovich, P. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics, 10*, 161.

Robinson, M., & Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics, 23*, 2881-2887.

Matukumalli, L., & Schroeder, S. (2009). Sequence based gene expression analysis. In D. Ed-wards, J. Stajich, & D. Hansen (Eds.), Bioinformatics, Tools and Applications (pp.191-207). New York, Springer.

Bradford, J., Hey, Y., ..., & Miller, C. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics, 11*, 282.

LaFranzo, N. (2013). Advantages of RNA-seq over Microarray Technology. [Online]. Available, https://cofactorgenomics.com/advantages-rna-seq-over-microarray-technology/

Malone, H., J., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol, 9*, 34. https://doi.org 10.1186/1741-7007-9-34

Wang, L., Feng, Z., ..., & Zhang, X. (2010). DEGseq: An R package for iden-tifying differentially expressed genes from RNA-seq data. *Bioinformatics., 26*(1), 136–138. https://doi.org/10.1093/bioinformatics/btp612

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). 'edgeR: A Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics, 26*, 139–140.

Hardcastle, T., & Kelly, K. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics, 11*, 442.

Manthey, A.,L., Terrell, A.,M., …, & Melinda K. Duncan, M.,K. (2014). Development of novel filtering criteria to analyze RNA-sequencing data obtained from the murine ocular lens during embryogenesis. *Genom Data.2014 Dec, 2,* 369–374. https://doi.org 10.1016/j.gdata.2014.10.015.

Balwierz, P., Carninci, P., ..., & van Nimwegen, E. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biology 10,* R79.

Oshlack, A., & Wakefield, M., J. (2009). Transcript length bias in RNA-seq data confounds systembiology. *Biology Direct, 4,* 14. https://doi.org/10.1186/1745-6150-4-14

Dillies, MA., Rau, A., ..., & Jaffrézic, F.(2013).A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform., 14*(6), 671-83. https://doi.org10.1093/bib/bbs046.

Robinson, M., & Smyth, G. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321-332.

Noble, S., W. (2009). How does multiple testing correction work? Nature Biotechnology27, 1135 – 1137. https://doi.org/10.1038/nbt1209-1135

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.Series B (Methodological). 57*(1), 289–300.

Storey, J., D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics., 31*(6), 2013–2035.

Storey, J., D. (2003). A direct approach to false discovery rate. *J R Stat SocSer B., 64* 479-498.

Anjum, A., Jaggi, S., V., …, & Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *J Comput Biol., 23*(4), 239–247.
https://doi.org10.1089/cmb.2015.0205

Kukurba, K., R., & Montgomery, S., B. (2015). RNA Sequencing and Analysis. *Cold Spring HarbProtoc.* 951–969. https://doi.org 10.1101/pdb.top084970

Tusher, V., G., Tibshirani, R., & Chu, G. (2000) .Significance analysis of microarrays applied to the ionizing radiation response. *Virginia Goss Tusher, Current Issue 98*(9), 5116–5121. https://doi.org10.1073/pnas.091062498

Storey, J., D., & Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proc Natl AcadSci U S A. 100*(16), 9440-5.

Smyth, G. K. (2004).Linear models, and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol. 3,* Article3.

Anders, S., & Huber, W. (2016). Differential expression of RNA-Seq data at the gene level -the DESeq package.

Zhang, Z. H., Jhaveri, D. J., ..., & Yi Zhao, Q. (2014, August 13). A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS One., 9*(8), e103207. https://doi.org10.1371/journal.pone.0103207.