OA Bioinformatics
Open Access

Genome Bioinformatics

Review

# Differential expression analysis methods for ribonucleic acid-sequencing data

AM Eteleeb*, EC Rouchka*

## Abstract

**Introduction**

High-throughput mRNA sequencing (also known as RNA-Seq) promises to be the technique of choice for studying transcriptome profiles, offering several advantages over old techniques such as microarrays. This technique provides the ability to develop precise methodologies for a variety of RNA-Seq applications, including gene expression quantification, novel transcript and exon discovery, differential expression (DE) analysis and splice variant detection. With the introduction of this technique, there has been a significant effort in developing new methods and statistical models to accurately model RNA-Seq data and test for differences in gene expression between biological conditions. In this review, we examine some of the most recently and widely used methods for DE analysis. We provide a detailed review of these methods by looking at the following three main aspects: statistical methods for normalisation, statistical modelling of gene expression and statistical methods for DE testing.

**Conclusion**

No single DE method can be considered as the best among available methods. Some methods perform well in particular situations, but their performance is poor in others.

## Introduction

With the advent of next generation sequencing (NGS) technologies, where a large volume of short deoxyribonucleic acid (DNA) sequences (reads) are generated, new methods and techniques have been developed for transcriptome analysis. Ribonucleic acid-sequencing (RNA-Seq) technology, which is based on the direct sequencing of complementary DNA (cDNA)[1], provides the ability for the reconstruction of transcripts, estimation of mRNA abundances and testing for differential expression (DE) genes between two or more conditions. This technology has enabled researchers and scientists to study the transcriptome at an unprecedented rate and has lately become a common platform for transcriptome analysis. This technique offers several advantages over the old microarray technology[2]. For instance, whereas microarrays generate expression signal intensities, RNA-Seq data generates digital gene expression counts. Unlike microarrays, RNA-Seq has a low background noise with high resolution. While microarrays offer resolution at the probe length, RNA-Seq allows for a single base resolution. Such granularity allows for the detection of splice variants. The dynamic range for quantifying expression differences is limited to a few hundred folds in microarrays, and can be nearly 10,000 fold with RNA-Seq data. One key limitation for microarrays is that they rely on a reference genome while RNA-Seq can take advantage of such an annotation. It also offers the ability for *de novo* transcriptomics.

An RNA-Seq experiment starts with the extraction of total RNA or a portion, such as polyadenylated-RNA[2]. The extracted RNA is then converted to a library of double-stranded cDNA and sheared into small fragments. In the next step, adapters are attached to one or both sides of each cDNA fragment. Using NGS platforms, such as Illumina's HighSeq 2500, Roche 454 GS FLX Sequencer, Applied Biosystems SOLiD Sequencer, Helicos HeliScope, or Pacific Biosciences/RS sequencer (Table 1 shows more detailed information about the most recently NGS platforms), each cDNA fragment is sequenced and a short sequence (read) from one end of the fragment (single-end tag) or from both ends (paired-end tag) is obtained (Figure 1). The obtained reads are mapped to the reference genome or transcriptome to measure the abundance of each transcript. If the reference genome or transcriptome is not available, short sequences (reads) can be assembled *de novo* to identify the full set of transcripts, followed by abundance estimation.

One of the primary applications in RNA-Seq is the study of gene expression profiling across experimental conditions. The number of reads that map to a gene is an approximation of its expression at the transcription level. Thus, the study of determining which genes have changed significantly in terms of their expression across biological samples is referred to as DE analysis. This step is essential in any RNA-Seq study. Identifying which genes are DEs between samples help researchers to understand the functions of genes in response to a given condition. In this review, we examine the most recently developed and widely used methods for DE analysis. We observe different statistical models that each method uses to test for DE. Because a large number of methods and tools have been developed in the last few years for DE analysis, not all DE methods

* Corresponding authors
Emails: ametel01@louisville.edu; eric.rouchka@louisville.edu

Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, USA



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of Interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

Review

| Table 1. Detailed information for current NGS platforms. | | | | |
|---|---|---|---|---|
| Platform | Amplification | Read length | Throughput | Technology |
| Roche/GS FLX Titanium XL+ | Emulsion PCR | Up to 1,000 bp | 700 Mbp/run | Pyrosequencing |
| Illumina/HiSeq 2500 | Bridge PCR | 2 × 100 bp | 600 Gbp/run | Sequencing by synthesis |
| ABI/SOLiD 5500xl | Emulsion PCR | 50-100 bp | >100 Gbp/run | Sequencing by ligation |
| Polonator/G.007 | Emulsion PCR | 26 bp | 8-10Gbp/run | Sequencing by ligation |
| Helicos/Heliscope | No | 35 (25-55) bp | 21-35 Gbp/run | Single molecule sequencing |
| Pacific BioSciences/RS | No | 1,000-10,000 bp | 13 Gb/run | SMRT |
| IonTorrent/Proton I Chip | Emulsion PCR | 100-400 bp | 10 Gb/run | Semiconductor-based pH sequencing |

NGS, next generation sequencing; PCR, polymerase chain reaction; SMRT, single molecule real time.

are discussed here, but instead we emphasise on the most widely used methods, including DEGSeq[3], edgeR[4], DESeq[5], baySeq[6] and Cuffdiff[7]. A comprehensive list of the DE methods can be found in Table 2.

**Background**

The detection, of which genes have significant DE across samples, requires the use of statistical hypothesis tests to model RNA-Seq count data. For any DE analysis, the following three components should be considered: normalisation of read counts, statistical modelling of gene expression and testing for DE[8].

**Normalisation**

In order to derive an accurate comparison within and between samples, normalisation is performed on read counts to adjust for sequencing depth variations and other systematic technical variations, which results in a comparable data across conditions. Thus, to discover significant changes in expression, studies have shown that normalisation is an essential step in the analysis of DE. Several normalisation techniques have been proposed in the published literature. Marioni et al.[9] used the total read count (TC) to normalise read counts. This normalisation method divides transcript read count by the total number of reads as follows: $\frac{X_{ij}}{N_j}$, where $X_{ij}$ is the number of reads for gene $i$ in sample $j$ and $N_j$ is the number of reads in sample $j$ (library size). Such an approach is equivalent to the total intensity normalisation procedure applied for microarrays. Bullard et al.[10] proposed a method similar to the TC method, quantile

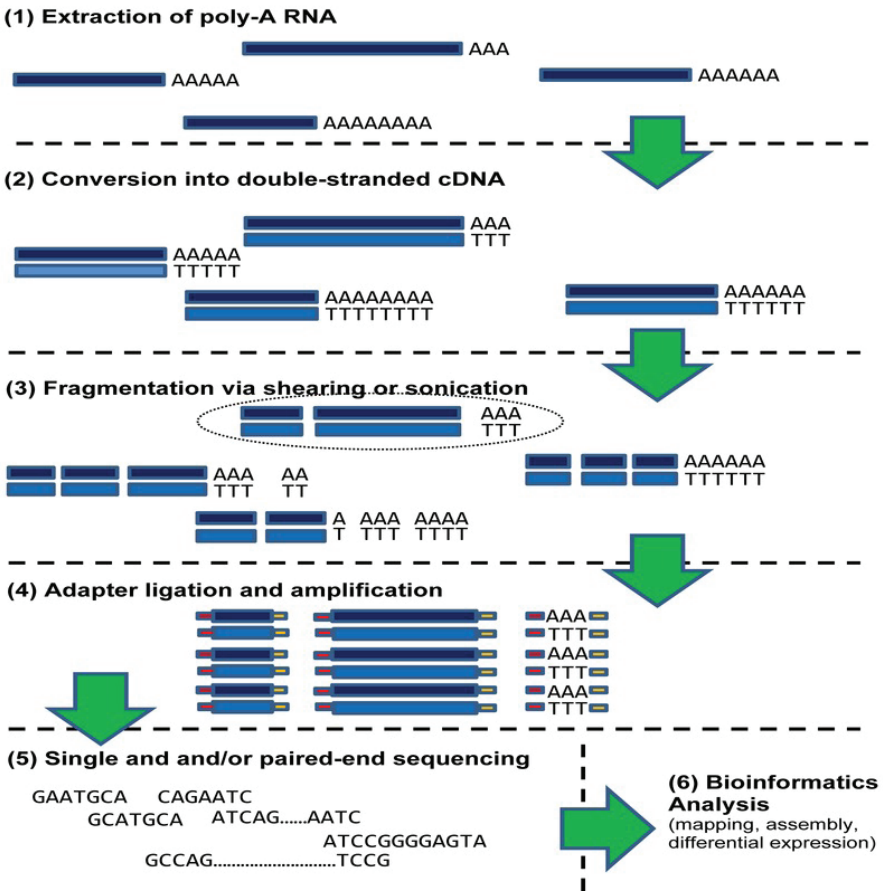


***Figure 1:*** The workflow of an RNA-Seq experiment.
cDNA, complementary deoxyribose nucleic acid; RNA, ribonucleic acid



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

Review

| Table 2. List of common differential expression methods. | |
|---|---|
| **Method** | **Technique** |
| DEGseq (Wang et al., 2010)[3] | MA-plots based method, assuming normal distribution for M \| A. |
| edgeR (Robinson et al., 2010)[4] | Exact test based on NB distribution. |
| DESeq (Anders et al., 2010)[5] | Exact test based on NB distribution. |
| baySeq (Hardcastle et al., 2010)[6] | Empirical Bayesian method (compute posterior probabilities of models, based on Poisson or NB distribution). |
| Cuffdiff (Trapnell et al., 2010)[7] | NB distribution to model the variance in fragment counts. |
| LRT (Marioni et al., 2008)[9] | Likelihood ratio test based on Poisson model. |
| PoissonSeq (Li et al., 2011)[16] | R package based on Poisson log-linear model. |
| GPseq (Srivastava et al., 2010)[18] | Likelihood ratio test for two-parameter generalised Poisson model. |
| NOISeq (Tarazona et al., 2011)[13] | Empirical approach to model the noise distribution of DE by contrasting fold-change differences (M) and absolute expression differences (D) for all the features in samples within the same condition. |
| EBSeq (Leng et al., 2012)[15] | Empirical Bayesian approach that models a number of features observed in RNA-Seq data. |
| SAMSeq (Li et al., 2011)[21] | Non-parametric approach for identifying differential expression in RNA-Seq data. |
| npSeq (Li et al., 2011)[21] | Non-parametric approach for identifying differential expression in RNA-Seq data. Similar to SAMSeq with only difference that npSeq uses symmetric cut-offs, while SAM uses asymmetric cut-offs. |
| NBPSeq (Di et al., 2011)[22] | NB models for two-group comparisons and regression inferences from RNA-sequencing data. |
| ShrinkSeq (Wiel et al., 2012)[23] | Bayes-empirical Bayes method that analyses RNA-Seq data by estimating multiple shrinkage priors. It supports a variety of count models such as NB mode. |
| TSPM (Auer et al., 2011)[24] | Two-Stage Poisson Model for testing RNA-Seq data. |
| Limma (Smyth et al., 2004)[14] | An R package that uses linear models for the analysis of gene expression data arising from microarray or RNA-Seq technologies. |
| Alexa-Seq (Griffith et al., 2010)[25] | A method to analyse massively parallel RNA sequence data to catalogue transcripts and assess differential and alternative expression of known and predicted mRNA isoforms in cells and tissues. |
| ASC (Wu et al., 2010)[26] | Empirical Bayes method to detect differential expression. |
| BBSeq (Zhou et al., 2011)[27] | A method designed for the DE analysis of the RNA-Seq count data. The method incorporates the following two approaches: (1) a simple beta-binomial generalised linear model and (2) mean-overdispersion model used to capture the gene specific dispersion. |
| DiffSplice (Hu et al., 2012)[28] | An *ab initio* method for the detection of alternative splicing isoforms that are DE under different conditions using RNA-Seq reads. DiffSplice does rely on annotated transcriptome or pre-determined splice pattern. |
| QuasiSeq (Lund et al., 2012)[29] | An R package used to apply the QL, QLShrink and QLSpline methods to quasi-Poisson or quasi-NB models for identifying DEGs in RNA-Seq data. |
| BitSeq (Glaus et al., 2012)[30] | A Bayesian approach for estimation of transcript expression level from RNA-Seq experiments and estimating DE between conditions. |
| MATS (Shen et al., 2011)[31] | Multivariate Analysis of Transcript Splicing (MATS) is a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. |
| Myrna (Langmead et al., 2010)[32] | A cloud computing tool for calculating differential gene expression in large RNA-Seq datasets. It includes short read alignment with interval calculations, normalisation, aggregation and statistical modelling. It uses both parametric and non-parametric tests. |
| CEDER (Wan et al., 2011)[33] | R/Bioconductor package developed to detect DEGs using RNA-Seq by combining significance of exons within a gene. |
| DEXSeq (Anders et al., 2012)[34] | R/Bioconductor package that finds differential exon usage based on RNA-Seq exon counts. It uses GLMs of the NB distribution (NB-GLMs) to model exon counts. |
| SplicingCompass (Aschoff et al., 2013)[35] | A method to predict genes that are differentially spliced between two different conditions using RNA-Seq data. It uses geometric angles between the high dimensional vectors of exon read counts. |
| MISO (Katz et al., 2010)[36] | Mixture of Isoforms (MISO) is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across samples. |

DE, differential expression; mRNA, mitochondrial ribonucleic acid; NB, negative binomial; RNA-Seq, ribonucleic acid sequencing.



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of Interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access



normalisation, borrowed from microarray technology, by replacing the total counts by the upper quantile (UQ) of the counts. The main concept of quantile normalisation is to match the distribution of read counts in each lane to a reference distribution defined in terms of median counts across sorted lanes. Replacing the UQ by the median, another form of quantile-based normalisation called median, is used. To correct for differences in library sizes and gene length, Mortazavi et al.[1] introduced reads per kilobase of transcript per million mapped reads (RPKM). The RPKM technique is defined as follows: $RKPM = 10^9 \frac{C_g}{l(t)N}$, where, $C_g$ is the number of reads mapped to gene $g$, $l(t)$ is the length of transcript $t$ in nucleotides, and $N$ is the total number of mappable reads in the sample. There are two cases in this context to consider. In the first case, when DE analysis is used to compare genes within a sample (each gene is compared relative to other genes in the sample), the length of the gene is important and should be considered for normalisation to avoid bias. This is clear because longer transcripts will by their nature have more read counts. In this case, read counts should be normalised by gene length, in effect reducing counts to the nucleotide base level. RPKM has been used widely to normalise read counts by using both the library size and the gene length. In the second case, when DE analysis is applied to compare the expression of the same genes in different samples, the gene length is not considered in the normalisation procedure. This is also clear because genes have the same lengths across samples. As an alternative to RPKM, the transcripts per million (TPM)[1], procedure normalises RNA-Seq data by dividing the number of reads of a transcript by the total clone count of the sample multiplied by 1,000,000. Results using this method are reported as reads/TPM for each sample. One of the limitations of TPM is the inability to handle datasets marked with different RNA composition. Thus, another method called trimmed mean of M-values (TMM) was proposed by Robinson et al.[11], as an attempt to remove RNA compositional bias. By estimating the relative RNA production levels, TMM equates the overall expression levels of genes between samples under the assumption that a large number of the genes are not differentially expressed. To calculate the normalisation scaling factor, this method uses a weighted trimmed mean of the log ratios between two samples[11].

DE methods use different normalisation procedures, some of which have improved the procedures mentioned above. For example, Marioni et al.[9] used the TC method to normalise counts; DEGSeq provides three choices for normalisation, 'none', 'median', and 'loess'; Mortazavi et al.[1] used RPKM and Trapnell et al.[7] implemented a slightly modified version of RPKM called fragments per kilobase of exon per million mapped fragments (FPKM) in their Cuffdiff method.

R Bioconductor packages, such as edgeR, DESeq, and baySeq use different techniques as well. While DESeq and baySeq use the library size for the normalisation procedure, edgeR implements the TMM method. DESeq uses the median of scaled counts (similar to the quantile normalisation) to estimate the normalisation[12]. For each sample, the DESeq scaling factor is computed for each gene as the median of the ratio of its read count over its geometric mean across all samples[5,8]. Using the assumption that most genes are not DE, DESeq uses the median of ratios associated with each sample to obtain the scaling factor. NOISeq, as proposed by Tarazona et al.[13], uses several options for normalisation, including TMM, RPKM and UQ. The Bioconductor package Linear Models for Microarray Data (limma)[14], an R package designed initially for DE analysis of microarray data, but lately is adapted for RNA-Seq data, implements quantile normalisation. EBSeq[15] provides two choices for normalisation, either by using the median of scaled counts used in DESeq or by using quantile normalisation approach. PoissonSeq[16] uses a normalisation procedure, which assumes a Poisson model for the data.

**Statistical modelling of gene expression**

The detection of which genes have changed significantly between biological samples, requires the use of statistical hypothesis tests, to model count data from RNA-Seq experiments. Currently, most statistical models are based on parametric assumptions for modelling RNA-Seq data. Discrete probability distributions, such as binomial, Poisson and negative binomial (NB) distributions, have been used to model RNA-Seq count data[12]. In RNA-Seq studies that use a single source of RNA, the distribution of counts across technical replicates for the majority of the gene was indeed Poisson[12,17], in the form of $f(n,y) = (\lambda^n e^{-\lambda})/n!$, where $n$ is the number of read counts and $\lambda$ is the expected number of reads in each transcript[8]. Early methods, such as the likelihood ratio test (LRT) proposed by Marioni et al., DEGSeq, PoissionSeq and Gpseq[18], have been developed to detect database of essential genes (DEGs) based on this distribution. However, the Poisson distribution suffers from the inability to capture biological variability within RNA-Seq data[12,17], because the variance of the Poisson distribution is equal to the mean. Because the variance of many genes is likely to exceed the mean, this results in over-dispersion. Thus, Poisson-based analyses using biological replicates will be prone to high false positive rates.

To address over-dispersion and account for biological variability, methods such as edgeR, DESeq, baySeq and Cuffdiff, have been developed based on the NB distribution to model read counts. These methods address over-dispersion by defining



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

Review

the relationship between the variance $v$ and mean $\mu$. For example, edgeR and DESeq define this relationship as $v = \mu + \alpha\mu^2$, where $\alpha$ is the dispersion factor. edgeR provides two options for $\alpha$, a common dispersion (estimated from all genes) and tag-wise dispersion (estimated for individual genes)[4,12,19]. DESeq, on the other hand, estimates the dispersion parameter by using a combination of two terms for the variance, one term estimates the Poisson (the mean expression $\mu$), and the second terms is the raw variance of the gene used to model the biological expression variability[5,8]. Cuffdiff computes two variance models, i.e., one for single-isoform genes and one for multi-isoform genes. For single-isoform genes, Cuffdiff computes the expression variance similar to DESeq using NB distribution. When a gene has multiple isoforms, Cuffdiff models over-dispersion by using the beta NB distribution[7]. BaySeq differs from the above three methods and implements an empirical Bayesian model based on NB distribution. This model estimates the prior probability parameters by bootstrapping from the data and then applies the maximum likelihood method. PoissonSeq models RNA-Seq count data by using a Poisson log-linear model. The mean $\mu_{ij}$ in this model is defined as a log-linear model $log\mu_{ij} = logd_i + log\beta_j + Y_jY_i$, where, $d_i$ is the library size of sample $i$, $\beta_j$ is the expression level of gene $j$, and $Y_j$ is the correlation of gene $j$ with condition $y_i$[8,9,16]. If there is no association between gene $j$ and $y_i$, then $Y_j = 0$ and $Y_j \neq 0$ otherwise.

**Testing for differential expression**

Once the parameters are estimated, statistical tests such as $t$-test, Wilcoxon test, or Fisher's exact test (FET), can be applied on the normalised data, to detect significant differentially expressed genes between samples. Both DESeq and edgeR use a variation of the FET adopted for an NB distribution. Cuffdiff compares the log ratio of gene expression in two conditions against the log ratio of one and calculates the test statistics as follows: $T = \frac{E[log(y)]}{Var[log(y)]}$, where $Y$, is the log ratio of the normalised counts between the two conditions $\left(Y = \frac{FPKM_a}{FPKM_b}\right)$. baySeq employs an empirical Bayesian approach to determine DE between these conditions. For every gene, baySeq estimates two models, one assumes that the expression pattern is the same and a second assumes that the expression pattern is different across conditions. Thus, the posterior likelihood can be estimated using the prior estimates and the likelihood of the distribution of the data to decide if a gene is differentially expressed. PoissonSeq tests for DE by determining the significance of the correlation term $Y_j$ in the linear model using a score statistic[8,16]. The p-value is then derived using a *chi*-square distribution, because the score statistic is shown to follow this distribution. Other DE methods use different statistical tests to test for DEGs. For example, limma uses a moderated $t$-statistic to derive the p-value.

**Methods**

*Cuffdiff*

Cuffdiff is a Cufflinks module that aims to find significant changes in transcript expression, splicing, coding output and promoter use. It uses the Cufflinks transcript quantification module to calculate transcript/gene expression levels and tests them for significant changes. The main input of Cuffdiff is the reference transcripts as a gene transfer format (GTF) file and two or more sequence alignment map (SAM) or binary version of SAM (BAM) files containing the fragment alignments for two or more samples. The output of Cuffdiff is a set of several files containing changes in expression at the level of isoforms, primary transcripts and genes. To test for DE, Cuffdiff compares the log ratio of gene expression in two conditions against the log ratio of one and calculates the test statistics. This ratio requires the knowledge of the variance of the expression level in each condition, which is calculated for a transcript's expression levels as follows:

$$Var[FPKM_t] = \left(\frac{10^9}{l(t)M}\right)(Var[X_t])$$

where, $Var[X_t]$ is the variance in the number of fragments coming from the transcripts across replicates. Cuffdiff uses the NB distribution to model the variance in fragment counts across replicates and the square root of the Jensen-Shannon (JS) divergence to quantify the changes in relative abundance. Thus, if we have abundances $p_1, p_2, \ldots, p_n$, then the *entropy* of the discrete distribution is defined as follows:

$$H(p) = -\sum_{i=1}^{n} p_i log p_i$$

and the JS divergence between a set of $m$ distributions $p^1, p^2, \ldots, p^m$ is defined as follows:

$$JS(p^{1,\ldots}, p^m) = H\left(\frac{p^{1,\ldots}, p^m}{m}\right) - \frac{\sum_{j=1}^{n} H(p^j)}{m}$$

Based on this JS divergence, Cuffdiff assigns p-values to the observed changes.

*edgeR*

The R Bioconductor package, edgeR, was initially developed for SAGE, but because the methods are applicable to RNA-Seq, it has been also used for detecting DE in RNA-Seq data. edgeR is based on the NB distribution if data are over-dispersed. However, in cases where there is no over-dispersion, the Poisson model is used. The edgeR count model is defined as follows: $Y_{gij} \sim NB\ (M_j p_{gi}, \varphi)$, where $Y_{gij}$ represents the observed data for gene $g$ in sample $j$ and experimental group $i$. The parameter $M_j$ denotes the total number of reads in a sample (library size), whereas the parameter $p_{gi}$ represents the relative abundance of gene $g$ in group $i$. $\varphi$ is the dispersion parameter. In the case of over-dispersion, the NB model is parameterised with the mean $\mu_{ig} = M_j p g_i$ and variance $v = \mu_{gi} + \mu^2_{gi}\varphi$. However, in the case of no over-dispersion ($\varphi = 0$), the NB model is reduced to Poisson model.



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of Interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

Review

The main input to edgeR is a table of counts constructed as a matrix, whose rows represent biological feature (e.g., genes, transcripts, or exons) and columns represents different samples. The output is a list of differentially expressed genes.

*DEGSeq*

DEGSeq is another R Bioconductor package developed for RNA-Seq data. The statistical model this package uses is based on a Poisson distribution. Two novel methods have been proposed in this package, an MA-plot-based method with random sampling and an MA-plot-based method with technical replicates, where M is the log ratio of the counts between two conditions for gene $g$ and $A$ is the average of the log concentration of the gene in the two groups[19]. Along with those two methods, three existing methods, FET, LRT and samWrapper, have been integrated into DEGSeq to identify differential expressed genes. In the MA random sampling, RNA sequencing can be modelled as a random sampling process, where each read is sampled independently and uniformly from every possible nucleotide in the sample. Thus, the number of reads coming from a gene/transcript follows a binomial distribution, which can be approximated by a Poisson distribution. With this assumption, DEGSeq is not applicable to data with over-dispersion, which limits its use for RNA-Seq analysis. The input of this package is uniquely mapped reads, a gene annotation of the corresponding genome, and gene expression counts for each sample. The output includes a text file, which contains the gene expression values for the samples, a P-value and two kinds of Q-values (adjusted p-values) and an extensible hypertext markup language (XHTML) summary page.

*DESeq*

DESeq is an R Bioconductor package that analyses RNA-Seq count data using the NB distribution and an estimator of the distribution's variance. DESeq uses a similar statistical model to edgeR, with a few extensions allowing for more general data-driven relationships of variance and mean. Under the assumption of a locally linear relationship between variance and mean expression levels, the variance can be estimated using data with similar expression levels[12]. The input of DESeq is a table of count data that reports for each sample, the number of reads that have been assigned to a gene. Thus, a table cell in the $i$-th and $j$-th column represents the number of reads mapped to gene $i$ in sample $j$. The output is a list of differentially expressed genes with p-values and q-values. The NB distribution that DESeq uses to model count data is defined as follows: Kij ~ ($\mu_{ij}$ $\sigma^2_{ij}$), where, $K_{ij}$ denotes the read counts for gene $i$ in sample $j$. This model has two parameters, the mean $\mu_{ij}$ and the variance $\sigma^2_{ij}$. These two parameters are often not known in advance and therefore, have to be estimated from the data. The mean $\mu_{ij}$ can be defined as follows: $\mu_{ij} = q_i,p(i)^S j$, which is the product of the expected read count (per gene and condition) $q_i,\rho(j)$ and size factor $S_j$, which represents the coverage of library $j$. $\rho(j)$ is the experimental condition of sample $j$. In contrast, the variance is defined as follows: $\sigma^2 = \mu_{ij} + s_j^2 v_i,\rho(j)$, where, $v_i,\rho(j)$, is the per gene raw variance parameter. This parameter is assumed to be a smooth function of $qi,\rho$ and defined as follows: $v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)})$, which should allow the pooling of data from genes with similar expression strength. To perform testing, DESeq uses FET on NB data. Thus, for two conditions A and B, the null hypothesis is that the counts of the two conditions are equal ($q_{iA}$=$q_{iB}$). The test statistic is performed using FET and the p-values computed using the following formula:

$$pi = \frac{\sum_{a+b=K_{iS}, p(a,b) \le p(K_{iA}, K_{iB})} p(a,b)}{\sum_{a+b=K_{iS}} p(a,b)}$$

where $k_{iA}$, and $k_{iB}$ are the TCs in each condition and $k_{iS} = k_{iA} + k_{iB}$. Variables $a$ and $b$ denote the even probabilities for any pair of numbers $a$ and $b$. For more details about the computation methods of the above models, refer to DESeq in the work of Anders and Huber[5].

*baySeq*

baySeq is an R Bioconductor package that assumes the data follows a NB distribution. baySeq differs from the above two packages in the strategy of estimating significance by employing an empirical Bayesian approach to determine DE across conditions. The baySeq approach starts by first bootstrapping to estimate prior parameters from the data and then assessing posterior likelihoods of the models by applying either maximum likelihood or quasi-likelihood methods[19]. In general, the baySeq approach aims to identify the behaviour of samples in terms of similarity and difference for each given model. Thus, for each gene, there will be two hypotheses, either the expression pattern is the same or different between two conditions. Under these two hypotheses, the posterior likelihood can be estimated using the prior estimates and the likelihood of the distribution of the data to decide if a gene is differentially expressed. The statistical models of baySeq are based on both Poisson and NB distributions.

The Poisson distribution is defined as follows: $Y_{gij} \sim (M_j p_{gi})$, assuming that the prior $p_{gi}$ follows a gamma distribution $p_{gi} = \Gamma(\alpha_{gi}, \beta_{gi})$. The second model, which is based on the NB distribution is defined as follows: $Y_{gij} \sim NB(M_j p_{gi}, \varphi_g)$. The baySeq package accepts the table of read counts (similar to DESeq, DEGSeq and edgeR) assigned to each gene for each sample as an input and reports a list of differentially expressed genes as an output.

## Discussion

In this review, some of the DE methods were examined. The focus



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access



of this review was on the most widely used methods, including edgeR, DESeq, DEGSeq, baySeq and Cuffdiff. We have looked at the following three main aspects: the normalisation procedure used by each method, the statistical model used for modelling RNA-Seq data and how each method tests for DE. In addition, we have given a brief description of the popular DE methods that are used by the majority of RNA-Seq community.

There is no an agreement in the published literature that a certain method is the optimal method under all circumstances. Thus, the reviewed methods perform differently under different conditions; hence, the choice of which particular method is to be used for a particular case depends on the experimental conditions[20]. Several comparison studies have been conducted in the literature evaluating DE methods to compare their performance under different conditions and using different datasets. Yet, there is no clear conclusion of those studies that certain methods are recommended to use for particular situations. Here, in this review, we tried to summarise the findings and observations of some of those studies in order to show the performance of the methods reviewed. Gao et al.[19] reviewed three R Bioconductor packages, namely, edgeR, DEGSeq, and baySeq. In their survey, they concluded that baySeq takes the longest time to run, DEGSeq is the easiest to use, but it does not handle over-dispersion data and edgeR is the most flexible package handling both Poisson and over-dispersion data[19].

Rapaport et al.[8] conducted a comprehensive evaluation of the commonly used DE methods using the Stroke Evaluation and Quality Committee (SEQC) benchmark data sets. Several features were evaluated in this study, including normalisation, accuracy of DE detection, modelling of gene expression and the effect of sequencing depth and number of replications on the process of DE detection. As a result of their evaluation, the authors found a significant difference between the methods with no single method, outperforming the others. However, by using different measurements, such as receiver operating characteristic (ROC) analysis and significant number of false positives, DE methods that model the data based on NB distribution (e.g., edgeR, DESeq and baySeq), have improved sensitivity and specificity. In addition, they were able to improve the control of false positive errors. This does not mean that those methods outperform the others. Methods, such as PoissonSeq and limma, which are not based on NB distribution, performed better as well. In some cases, PoissonSeq was just as good as edgeR and DESeq. Cuffdiff on the other hand did not perform as well as other methods, and has reduced sensitivity and specificity. Furthermore, they also found that limma, which was developed for microarray analysis, had a comparable performance.

Another comparison study was done by Kvam et al[12]. In this study, seven DE methods (not including Cuffdiff and PoissonSeq), were compared using a variety of simulations generated, based on different distribution models and real data. Using ROC curves and area under the ROC curves (AUC) as measures, Kvam et al. found that baySeq performs the best among others and ranks the highest number of truly DE genes. edgeR and DESeq perform similarly and close to baySeq, which is expected, because both methods use similar statistical models (based on NB) to model count data.

Soneson et al.[20] conducted another comparison between 11 DE methods (not including Cuffdiff and PoissonSeq). All compared methods in this study are R Bioconductor packages, which take the table of counts described above as an input. These 11 methods were evaluated using simulated and real data. The main focus in this study was on the following three aspects: the performance of ranking truly DE genes, type I error control and FDR and the computational time requirement. The AUC was used as a measure to evaluate the performance of each method. They concluded that all compared methods perform similarly when large sample sizes are used with a strong dependency on sample size for TSPM, EBSeq, SAMseq and baySeq. With the smallest sample sizes, DESeq, edgeR, NBPSeq and limma, performed the best among other methods. For controlling type I error, six methods which generate p-values were evaluated. The performance of the six methods was high with the note that DESeq was the most conservative among the six methods. TSPM and NBPSeq found the highest number of false positives. From their conclusion, limma performed well under many conditions and was not affected by outliers. In addition, it was computationally efficient. However, it has the limitation of requiring at least three replicates per sample to provide good results. SAMseq performed well with large samples, but like limma it requires at least 4–5 replicates per sample to have a good performance. TSPM as mentioned above was the most affected method by sample size. Because edgeR, DESeq, and NBPSeq uses similar statistical models to model count data, their performance was similar in terms of accuracy. Table 2 in Soneson et al.[20] summarises the findings and observations of this study.

As a process of evaluating and testing, our newly developed method, Island-Based (IB), we compared its performance to three DE methods, namely, Cuffdiff, DESeq and edgeR using two benchmarks MicroArray Quality Control (MAQC) RNA-Seq datasets. By using the AUC as a measure, our approach outperforms other methods in both the datasets. The performance of Cuffdiff was better than DESeq and edgeR, (both performed similarly) but not as well as our approach (Figure 2).



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of Interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

*Review*

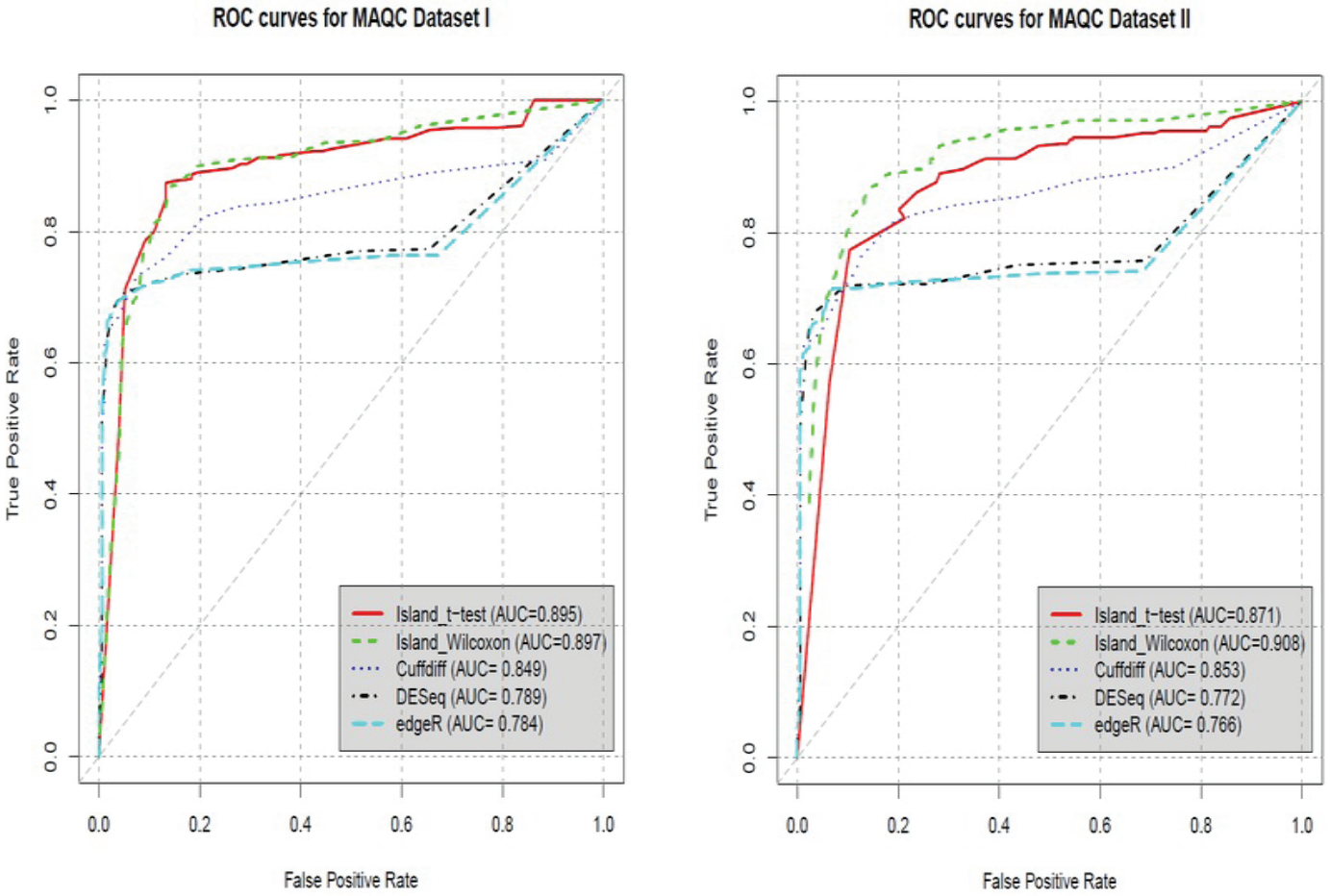


***Figure 2:*** The ROC curves for the four methods using qRT-PCR validated gene set for MAQC dataset I and II. MAQC, MicroArray Quality Control; PCR, polymerase chain reaction; ROC, receiver operating characteristic.

## *Conclusion*

From the above discussion, we can conclude that no single DE method can be considered as the best among the available methods. Some methods perform well in particular situations, but their performance is poor in other situations. While some methods require a number of biological replicates for each condition to perform well, others are not conservative and they work with/without replicates. There are several factors that affect the performance of detecting which genes are differentially expressed between samples. Examples of those factors are the normalisation procedure and the statistical model used. DE methods in general, differ mainly in those two factors, which make them provide different results and generate different lists of DE genes.

## *Abbreviations list*

AUC, area under the ROC curve; cDNA, complementary DNA; DE, differential expression; DEG, database of essential genes; DNA, deoxyribonucleic acid; FET, Fisher's exact test; limma, Linear Models for Microarray Data; LRT, likelihood ratio test; JS, Jensen-Shannon; NB, negative binomial; NGS, next generation sequencing; ROC, receiver operating characteristic; RNA-Seq, ribonucleic acid sequencing; RPKM, reads per kilobase of transcript per million mapped reads; Sam, sequence alignment map; TC, total read count; TMM, trimmed mean of M-values; TPM, transcripts per million; UQ, upper quantile.

## *Acknowledgments*

Funding was provided in part by the National Institutes of Health (NIH) grants 5P20GM103436-13, P20RR016481 (ECR) and 3P20RR016481- 09S1 (ECR, AME). The article contents are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health.

## *References*

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621–8.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.
3. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010 Jan;26(1):136–8.
4. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Bioinformatics
Open Access

*Review*

differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan;26(1):139–40.
5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Oct;11(10):R106.
6. HardcastleTJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug;11(1):422.
7. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5): 511–5.
8. Rapaport F, Khanin R, Liang Y, Krek A, Zumbo P, Mason CE, et al. Comprehensive evaluation of differential expression analysis methods for RNA-seq data. 2013 Jan; arXiv preprint arXiv:1301.5277.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008 Sep;18(9):1509–17.
10. Bullard JH, Purdom E, Hansen KD, Dudoit D. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010 Feb;11(1):94.
11. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010 Mar;11(3):R25.
12. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot. 2012 Feb;99(2):248–56.
13. Tarazona S, García-Alcalde F, Ferrer A, Dopazo J, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011 Dec;21(12):2213–23.
14. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004 Feb;3(1):3.
15. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013 Apr;29(8):1035–43.
16. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012 Jul;13(3):523–38.
17. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010 Dec;11(12):220.
18. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. 2010 Sep;38(17):e170.
19. Gao D, Kim J, Kim H, Phang TL, Selby H, Tan AC, et al. A survey of statistical software for analysing RNA-seq data. Hum Genomics. 2010 Oct;5(1):56–60.
20. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013 Mar;14(1):91.
21. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2011 Nov.
22. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol, 2011 May;10(1):24.
23. Van De Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013 Jan;14(1): 113–28.
24. Auer PL, Doerge RW. A two-stage Poisson model for testing RNA-Seq data. Stat Appl Genetics Mol Biol. 2011 May;10(1):26.
25. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, et al. Alternative expression analysis by RNA sequencing. Nat Methods. 2010 Oct;7(10):843–7.
26. Wu Z, Jenkins BD, Rynearson TA, Dyhrman ST, Saito MA, Mercier M, et al. Empirical Bayes analysis of sequencing-based transcriptional profiling without replicates. BMC Bioinformatics. 2010 Nov;11(1):564.
27. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011 Oct;27(19):2672–8.
28. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res. 2013 Jan;41(2):e39.
29. Lund SP, Nettleton D, McCarthy DJ, Smyth GK. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat Appl Genet Mol Biol. 2012 Oct;11(5):8.
30. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012 Jul;28(13):1721–8.
31. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, et al. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. 2012 Apr;40(8):e61.
32. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 2010 Aug;11(8):R83.
33. Wan L, Sun F. CEDER: accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq. IEEE/ACM Trans Comput Biol Bioinform. 2012 Sep–Oct;9(5):1281–92.
34. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012 Oct;22(10):2008–17.
35. Aschoff M, Hotz-Wagenblatt A, Glatting KH, Fischer M, Eils R, König R. SplicingCompass: differential splicing detection using RNA-seq data. Bioinformatics. 2013 May;29(9):1141–8.
36. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010 Dec;7(12):1009–15.



FOR CITATION PURPOSES: *Eteleeb AM, Rouchka EC*. Differential expression analysis methods for ribonucleic acid-sequencing data. OA Bioinformatics 2013 Sep 01;1(1):3.

*Competing interests: none declared. Conflict of Interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*