Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software

Huei-Chung Huang*, Yi Niu*,† and Li-Xuan Qin

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. *Both authors made equal contributions to the paper. †Current address: School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, China.

Supplementary Issue: Sequencing Platform Modeling and Analysis

ABSTRACT: Deep sequencing has recently emerged as a powerful alternative to microarrays for the high-throughput profiling of gene expression. In order to account for the discrete nature of RNA sequencing data, new statistical methods and computational tools have been developed for the analysis of differential expression to identify genes that are relevant to a disease such as cancer. In this paper, it is thus timely to provide an overview of these analysis methods and tools. For readers with statistical background, we also review the parameter estimation algorithms and hypothesis testing strategies used in these methods.

KEYWORDS: RNA sequencing, differential expression analysis, overview, statistical methods, software.

SUPPLEMENT: Sequencing Platform Modeling and Analysis

CITATION: Huang et al. Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. *Cancer Informatics* 2015:14(S1) 57–67 doi: 10.4137/CIN.S21631.

TYPE: Review

RECEIVED: May 13, 2015. **RESUBMITTED:** August 24, 2015. **ACCEPTED FOR PUBLICATION:** August 26, 2015.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Twelve peer reviewers contributed to the peer review report. Reviewers' reports totaled 3,896 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by NIH grants CA008748 and CA151947 (HCH, YN, and LXQ). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: qinl@mskcc.org

Introduction

In the past decade, deep sequencing has emerged as a powerful alternative to microarrays for the high-throughput profiling of gene expression. Comparing with microarrays, RNA sequencing (RNA-seq) possesses a number of technological advantages such as a wider dynamic range and the freedom from predesigned probes.[1–3] It also comes with a unique data feature as discrete sequencing reads. In order to account for this unique data feature, statistical methodologies and computational algorithms have been developed based on various data distributional assumptions such as Poisson, negative binomial, beta binomial, (full or empirical) Bayesian, and nonparametric.[4–16,]

For researchers who are new to the analysis of RNA-seq data, in this paper we provide an introductory overview of the methods and software available for the differential expression analysis (DEA) of RNA-seq data when the analysis goal is to identify genes that are relevant to a disease such as cancer.[1,17,18] In addition, for those who are interested in the statistical aspects of these methods, we also provide an overview of their parameter estimation algorithms and hypothesis testing strategies. The overview of these statistical aspects in our paper provides a unique contribution to the review literature on RNA-seq DEA methods.[3,18–23] For readers who are interested in a performance comparison of RNA-seq DEA methods, they can refer to a large body of such papers in the literature.[20–23]

The rest of the paper is organized as follows. In the Notation and Normalization Methods section, we introduce the unified notations used for the methods reviewed in our paper and touch on the normalization methods typically used to preprocess RNA-seq data before DEA. In the Statistical Modeling of RNA-seq Data section, we review the statistical modeling RNA-seq DEA categorized by the distributional assumptions such as Poisson,[4–6] negative binomial,[7–10] beta binomial,[11,12] Bayesian,[13,14] and nonparametric.[15,16] All reviewed methods directly work with gene-level count data for DEA and have available R packages. For interested readers with advanced statistical knowledge, the parameter estimation algorithm for each method is presented separately in a text box, and the typical statistical testing frameworks that have been proposed for RNA-seq DEA are reviewed in the Statistical Testing section. Finally, computational tools implemented for the

reviewed methods are summarized in Table 2. We note that the methods reviewed in this paper are not an exhaustive collection of available methods in the literature. Rather, we reviewed a list of most commonly used categories of modeling assumptions and included a few representative methods for each category, to help researchers who are new to the field orientated and started in the still evolving literature on this topic.

Notation and Normalization Methods

Notation. RNA-seq data for G genes and N samples can be described by a $G \times N$ matrix $\mathbf{Y}$. Each entry y_{gi} ($g = 1, \ldots, G$, $i = 1, \ldots, N$) represents the count of sequencing reads for gene g in sample i. For a given g and i, y_{gi} is a nonnegative integer representing the number of reads mapped to gene g in sample i. For succinctness, we also use notations "$\cdot$" for summations, eg, $y_{g\cdot} = \sum_{i=1}^{N} y_{gi}$ and $y_{\cdot i} = \sum_{g=1}^{G} y_{gi}$.

We use $\mathbf{X}$ to represent an $N \times P$ design matrix, where P is the number of covariates. For instance, x_{ip} can be an indicator variable of disease status, taking a value of 0 for a normal sample and a value of 1 for a tumor sample. When comparing K groups of samples, C_k represents the collection of indices of the samples in group k ($k = 1, \ldots, K$), that is, $C_k = \{i: x_i = k\}$. Each sample can only belong to one group.

Normalization methods. Similar to microarray data, RNA-seq data are also prone to nonbiological effects due to the experimental process. Consequently, these effects need to be adjusted before any further data analysis.[24] One major source of nonbiological effects is sequencing depth, which can be adjusted by rescaling the sequencing counts with factors that mimic sequencing depth.[25] Reads per kilobase per million reads (RPKM) is a simple adjustment that considers gene counts standardized by the gene length and the total number of reads in each library as expression values.[17,26] More sophisticated adjustment factors, including trimmed mean of M-values (TMM),[27] DESeq size factor,[28] and quantile-based normalizations such as upper quartile normalization,[18] are given in Table 1. Other sources of nonbiological effects for RNA-seq include gene length and GC-content,[21,29] whose effects are typically assumed to be consistent across samples for a given gene and hence cancel out in the analysis of differential expression. Interested readers can look up available normalization methods adjusting for gene length and GC-content in the publications such as Risso et al.[29], Benjamini and Speed,[30] and Hansen et al.[31]

Table 1. List of sequencing depth normalization methods and reference papers.

METHODS	RELEVANT REFERENCES
RPKM	Mortazavi et al.[26]
Upper-quartile, Median	Bullard et al.[18]
TMM	Robinson et al.[7]
DESeq	Anders and Huber[8]
Quantile	Bolstad et al.[32]

Statistical Modeling of RNA-seq Data

Poisson. *Overview*. Models for read counts originated from the idea that each read is sampled independently from a pool of reads and hence the number of reads for a given gene follows a binomial distribution, which can be approximated by a Poisson distribution. Based on the Poisson model assumption for repeated sequencings of a sample, Marioni et al.[17] proposed to use a log-linear model to model the mean difference between two samples and adopted the classical likelihood ratio test for calculating the P-values. Based on the same Poisson assumption, Bullard et al.[18] proposed to use two other test statistics, exact test statistics and score test statistics, in the generalized linear model (GLM) framework. Li et al.[6] proposed a method called *Poisson-Seq*, which adapts a two-step procedure for fitting a Poisson model. The method first estimates sequencing depths using a Poisson goodness-of-fit statistic and then calculates a score statistic based on a log-linear model. In addition, Wang et al.[4] developed an R package, *DEGseq*, to identify differentially expressed (DE) genes with an MA-plot-based approach. Langmead et al.[5] incorporated cloud computing in their method called *Myrna*.

Modeling. In a Poisson model, one assumes that Y_{gi}, the number of reads mapped to gene g in sample i, follows a Poisson distribution, $y_{gi} \sim \text{Poisson}(\mu_{gi})$. μ_{gi} is the rate parameter for gene g in sample i, which equals both the mean and the variance of the read counts. The probability mass function is:

$$f(y_{gi}|\mu_{gi}) = P(Y_{gi} = y_{gi}|\mu_{gi}) = \frac{\mu_{gi}^{y_{gi}} exp(-\mu_{gi})}{y_{gi}!} \qquad (3.1.1)$$

and $E(Y_{gi}) = \mu_{gi}$ and $\text{Var}(Y_{gi}) = \mu_{gi}$. The association of μ_{gi} with the same sample group can be described by a log-linear model as follows:

$$\log(\mu_{gi}) = \log d_i + \log \beta_g + \sum_{k=1}^{K} \gamma_{gk} I(i \in C_k), \qquad (3.1.2)$$

where d_i represents the sequencing depth of sample i and $\sum_{i=1}^{N} d_i = 1$ is assumed for generality. Let β_g be the expression level of gene g and γ_g be the association of gene g with the covariate. For hypothesis testing, $\gamma_{g1} = \ldots = \gamma_{gK} = 0$ indicates that the expression of gene g is not associated with the sample group. In the case of two sample group comparison, if $\gamma_g = 0$, then gene g is not DE between the two sample groups.

Algorithm Overview 1: Li et al.'s[6] *PoissonSeq*

Li and others proposed *PoissonSeq* that assumes the hypotheses as follows. Under the null hypothesis where genes and covariates are not relevant,

$$\log \mu_{gi} = \log d_i + \log \beta_g, \tag{3.1.a}$$

where d_i is the sequencing depth in sample i and β_g is the expression of gene g. The model fit from Equation (3.1.a) is denoted as $N_{gi}^{(0)}$ in later equations: $N_{gi}^{(0)} = \exp\left(\log\left(\hat{d}_i\right) + \log\left(\hat{\beta}_g\right)\right)$.

Under the alternative hypothesis where genes and covariates, x_i^*, are relevant,

$$\log \mu_{gi} = \log d_i + \log \beta_g + \gamma_g x_i^* \tag{3.1.b}$$

where x_i^* would be $I_{(i \in C_k)}$ when comparing two or multiple sample groups. The authors suggested using the maximum likelihood to estimate $\hat{\beta}_g$, as a result $\hat{\beta}_g = y_{g.}$. However, instead of using the maximum likelihood estimate of the sequencing depth in sample i, the authors sought for a set of genes, denoted by S, that are not DE to estimate sequencing depth in sample i:

$$\hat{d}_i = \frac{\sum_{g \in S} y_{gi}}{\sum_{g \in S} y_{g.}}. \tag{3.1.c}$$

They then estimated which genes belong to S by a Poisson goodness-of-fit statistic, ie,

$$\mathrm{GOF}_g = \sum_{i=1}^{N} \frac{\left(y_{gi} - \hat{d}_i y_{g.}\right)^2}{\hat{d}_i y_{g.}}. \tag{3.1.d}$$

S is set to be the genes whose GOF_g values are in the $(\varepsilon, 1-\varepsilon)$ quantile of all GOF_g values. Li and others used $\varepsilon = 0.25$ in their study.[6]

The objective is to test H_0:

$$\gamma_{g1} = \ldots = \gamma_{gK} = 0,$$

and score statistics were proposed to perform the testing. For a two-group or multiple-group covariate, the score statistic for gene g is

$$\sum_{k=1}^{K} \frac{\left[\sum_{i \in C_k}\left(y_{gi} - N_{gi}^{(0)}\right)\right]^2}{\sum_{i \in C_k} N_{gi}^{(0)}} \sim \chi^2(K-1). \tag{3.1.e}$$

With accumulating empirical data (especially with the data available for groups of multiple biological samples), researchers began to observe that in a group, the between-sample variation of sequencing reads for a gene often exceeds the mean.[17,23,33] This excessive variation that cannot be explained by the Poisson model is called overdispersion. Extensions of the classic Poisson model have been proposed in order to accommodate such overdispersion, including the two-stage Poisson models[34] and the generalized Poisson model.[35]

Negative binomial. *Overview*. A class of models based on the negative binomial distribution assumption has been developed in order to accommodate the overdispersion among biological replicate data.[8,9,33,36] Robinson and Smyth[33] used the conditional maximum likelihood (CML) to estimate the dispersion parameter–a measure of the excessive variance that a Poisson model does not incorporate–when assuming a common dispersion parameter across genes. They compared the CML method with alternative estimation methods based on pseudolikelihood, quasi-likelihood, and conditional inference.[37–39] In a follow-up paper,[36] they also extended the model to allow for gene-specific dispersion parameters and proposed to estimate the dispersion parameters by maximizing a weighted conditional likelihood with empirical Bayesian approximation. Details of their method, *edgeR*, can be found in Robinson and Smyth.[33,36] *edgeRun* is based on the same model as *edgeR* but it uses an unconditional exact test to achieve more power while paying the price of computational time.[40] Anders and Huber[8] proposed a method called *DESeq* also under the negative binomial assumption. They advocated the use of a robust estimate of normalization factors for the estimation of dispersion parameter and a local regression to obtain smooth function for each group on the graphs of expected proportions vs sample variances. *DESeq2* was developed in the study by Love et al.[9] as a successor of *DESeq*. It employs a number of new modeling features, such as the use of a shrunken fold change and a shrunken dispersion estimation method, to further improve the model performance. Di and others[10] proposed a method, *NBPSeq*, using a negative binomial power distribution instead of a regular negative binomial distribution. They hypothesized that $E\left(Y_{gi}\right) = \mu_{gi}, Var\left(Y_{gi}\right) = \mu_{gi}\left(1 + \phi\mu_{gi}^{\alpha-1}\right)$, and ϕ is common across genes while α helps to accommodate the overdispersion. ϕ and α are estimated by maximizing conditional log-likelihood,[41] conditional on the total gene counts for each gene g. An exact test modified for negative binomial power distribution is used for hypothesis testing. More details can be found in the study by Di et al.[10]

Modeling. The model setup for negative binomial is to assume $y_{gi} \sim$ negative binomial (μ_{gi}, ϕ_g). The dispersion parameter, ϕ_g, accounts for the sample-to-sample variability, which is usually assumed to be common across samples. There are various estimation methods for this model assumption. More specifically, the negative binomial probability mass function is written as

$$f(y_{gi}|\mu_{gi},\phi_g) = P(Y_{gi} = y_{gi}|\mu_{gi},\phi_g)$$
$$= \frac{\Gamma\left(y_{gi}+\phi_g^{-1}\right)}{\Gamma\left(\phi_g^{-1}\right)\Gamma\left(y_{gi}+1\right)}\left(\frac{1}{1+\mu_{gi}\phi_g}\right)^{\phi_g^{-1}}\left(\frac{\mu_{gi}}{\phi_g^{-1}+\mu_{gi}}\right)^{y_{gi}}, \quad (3.2.1)$$

where $E(Y_{gi}) = \mu_{gi}$ and $Var\left(Y_{gi}\right) = \mu_{gi} + \phi_g\mu_{gi}^2$. Hypothesis testing is set up as H_0: no difference either between the expected normalized expression of gene g in groups or between the proportion of reads that are gene g in groups.

Algorithm Overview 2: Overdispersion

Negative binomial can be derived as a gamma–Poisson mixture model (subscripts g's and i's are omitted for brevity), under the assumption that technical replicates follow a Poisson distribution, and biological replicates follow a gamma distribution, with the latter accommodating the overdispersion observed in empirical data.

$$y \sim \text{Poisson}(\mu), \mu \sim gamma(\alpha,\beta)$$
$$P(y|\mu) = \frac{\mu^y \exp(-\mu)}{y!}$$
$$f(\mu) = (\Gamma(\alpha)\beta^\alpha)^{-1}(\mu^{\alpha-1}\exp(-\mu/\beta))$$

Then,

$$P(y) = \int_0^\infty P(y_{gi}|\mu)f(\mu)d\mu$$
$$= (y!\Gamma(\alpha)\beta^\alpha)^{-1}\int_0^\infty \mu^{(y-\alpha)-1}\exp(-\mu(1+1/\beta))d\mu$$
$$= \frac{\Gamma(y+\alpha)\beta^y}{y!\Gamma(\alpha)(1+\beta)^{y+\alpha}}$$

One substitutes back μ_{gi}, y_{gi}, $\alpha = \phi_g^{-1}$, and $\beta = \mu_{gi}\phi_g$, a gamma–Poisson mixture can be viewed as a negative binomial, see Equation (3.2.1).

Algorithm Overview 3: Robinson and Smyth's[33,36] *edgeR*

In *edgeR*, $\mu_{gi} = m_i\lambda_{gk(i)}$ where m_i is the ith library size and $\lambda_{gk(i)} = \sum_{i=1}^{C_k}\lambda_{gi}$ represents the proportion of the total reads that is gene g in group k and λ_{gi} is the proportion of the total reads that is gene g in sample i.

Under the assumption of gene-wise (or tag-wise in the original paper) dispersion, ϕ_g is estimated by maximizing a weighted conditional log-likelihood, $WL(\phi_g)$:

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g) \quad (3.2.a)$$

where α is the weight given to the common likelihood, l_C; the maximum estimator of $WL(\phi_g)$ is denoted by $\hat{\phi}_g^{WL}$. An α has to be chosen such that $\hat{\phi}_g^{WL}$ coincides with an empirical Bayesian solution, $\hat{\phi}_g^B$, the Bayesian posterior mean estimator of ϕ_g where $\hat{\phi}_g|\phi \sim N(\phi_g,\tau_g^2)$ and $\phi_g \sim N(\phi_0,\tau_0^2)$ for $g = 1, \ldots, G$. The approximation method is selected as a direct estimate of ϕ_g is difficult because of the lack of a conjugate prior for ϕ in negative binomial model. Details are given in the study by Robinson and Smyth.[33]

In the study by Robinson and Smyth,[36] the overdispersion parameter is assumed to be common across all genes (ie, $\phi_g = \phi$). To estimate the shared dispersion parameter with and without equal library size, the authors proposed to use the CML and quantile-adjusted CML (qCML) as follows.

In a special case where $m_i = m$ for $i \in C_k$ where $C_k = \{i: k(i) = k\}$, $y_{gi} \sim$ negative binomial $(\mu_{gi} = m\lambda_{gk}, \phi)$ in group k and Y_{gi}'s evidently become identically distributed, and the maximum likelihood estimator (MLE) $\hat{\lambda}_{gk(i)}$ becomes $\frac{\sum_{i\in C_k} y_{gi}}{\sum_{i\in C_k} m_i}$ in group k. CML function for dispersion ϕ given $z_k = \sum_{i=1}^{n_k} y_{ki}$ was proposed. The function is as follows:

$$l_C(\phi) = \sum_{g=1}^{G} l_g(\phi) = \sum_{g=1}^{G}\sum_{k=1}^{K}[\sum_{j=1}^{n_k}\log\Gamma(y_{ki}+\phi^{-1}) + \log\Gamma(n_k\phi^{-1}) - \log\Gamma(z_k+n_k\phi^{-1}) - n_k\log\Gamma(\phi^{-1})] \quad (3.2.b)$$

In the case of different m_i in group k, the MLE of $\lambda_{gk(i)}$ depends on ϕ (ie, maximum likelihood estimation of the two parameters proceeds jointly). As a result, an approximate approach called qCML was proposed to equate the library sizes. The quantile-adjusted pseudo-data supposedly allows one to use a common likelihood $l_C(\phi)$ to estimate an accurate estimate of ϕ. Specifically, let $m^* = (\Pi_{i=1}^N m_i)^{\frac{1}{N}}$, where m^* is the geometric mean of the library sizes. Then, the observed data could be adjusted as if they were all sampled as identically distributed negative binomial $(m^*\lambda, \phi)$.

Hypothesis testing is set up as H_0: $\lambda_{g1} = \lambda_{g2}$, in other words, no difference in proportion of gene g in samples between group 1 and group 2.

Algorithm Overview 4: Anders and Huber's[8] *DESeq*

The read count y_{gi} is modeled by a GLM of negative binomial distribution with a log link:

$$\log(\lambda_{gi}) = \sum_{p=1}^{P} x_{ip}\beta_{gp} \quad (3.2.c)$$

The mean μ_{gi} is the proportion of reads for gene g in sample i, $\lambda_{gk(i)}$, scaled by a normalization factor, m_i. The variance σ^2_{gi} is $\mu_{gi} + m_i^2 v_{gk(i)}$, where $v_{gk(i)}$ is assumed to be a per gene raw variance, a smoothing function of λ_g and k. The use of the smoothing function can help stabilize the variance estimates especially when the number of samples is small. For the estimation of the normalization factor (which is referred to as the size factor by Anders and Huber), m_i, for each sample, the authors noted that highly DE genes are more likely to be influential on total count and so the median of the ratios of counts should be used for more robustness:

$$\hat{m}_i = \underset{i}{\text{median}} \frac{y_{gi}}{\left(\prod_{v=1}^{N} y_{gv}\right)^{1/N}} \quad (3.2.d)$$

Since $\lambda_{gk(i)}$ is proportional to the expected value of the unknown proportion from gene g in group k, it is estimated by the average of counts from all samples in group k with a common scale.

$$\hat{\lambda}_{gk(i)} = \frac{1}{M_k} \sum_{i:k(i)=k} \frac{y_{gi}}{\hat{m}_i}, \quad (3.2.e)$$

where M_k is the total number of replicates for group k. The sample variances with the common scale are calculated as:

$$w_{gk} = \frac{1}{M_k - 1} \sum_{i:k(i)=k} \left(\frac{y_{gi}}{\hat{m}_i} - \hat{\lambda}_{gk(i)} \right)^2 \quad (3.2.f)$$

$$z_{gk} = \frac{\hat{\lambda}_{gk(i)}}{M_k} \sum_{i:k(i)=k} \frac{1}{\hat{m}_i} \quad (3.2.g)$$

In the case of a sufficiently large number of M_k, one can see $w_{gk} - z_{gk}$ as the unbiased estimator of the raw variance v_{gk}. In the case of a small number of M_k, local regression for a smooth function $w_k(\lambda)$ on the graph of $(\hat{\lambda}_{gk(i)}, w_{gk})$ was suggested so that $w_k(\hat{\lambda}_{gk(i)}) - z_{gk}$ would be the estimate for the raw variance. More details are in the study by Anders and Huber.[8]

Algorithm Overview 5: Love et al.'s[9] *DESeq2*

DESeq2 allows the normalization factors to be gene specific (m_{gi}), rather than being fixed across genes (m_i). The estimation of m_{gi} is implemented in their new R packages.[9]

When modeling dispersion parameters, a large variation in estimates usually arises because of small sample sizes. *DESeq2* proposed to pool genes with similar average expression together for the estimation of dispersions. To do this, one first separately estimates dispersion with maximum likelihood. Then, one identifies a location parameter for the distribution of the estimates by fitting a smooth curve dependent on average normalized expressions, before finally shrinking gene-specific dispersions to the fitted curve using an empirical Bayesian approach. The authors stated that this procedure is more superior than DESeq.

In order to avoid identifying differential expressions in genes of small average expression, fold change estimation is shrunken toward 0 for genes with insufficient information by employing an empirical Bayesian shrinkage. The procedure is as follows: (1) obtain the maximum likelihood estimates for the log fold changes from the GLM fit, then (2) fit a normal distribution with mean 0 to the estimates, and (3) use that as the prior for a second GLM fit. The maximum a posterior and the standard error for each estimate are the products of this procedure and will be used for the calculation of Wald statistics for DEA.

DESeq2 computes a threshold, η, to filter genes based on their average normalized expressions. The threshold is calculated for maximizing the number of genes with a user-defined false discovery rate. The authors claimed that this filtering step effectively controls the power of detecting DE genes. The null hypothesis becomes $|\beta_{gp}| \leq \eta$ where β_{gp} is the shrunken log fold change.

Finally, the method provides a way to diagnose outliers using the Cook's distance from the GLM within each gene, C_d. Samples are flagged with $C_d \geq 99\%$ quantile of an F distribution with degrees of freedom as the number of parameter, P, and the difference in the number of samples and the number of parameter, $N - P$. When there is a large number of replicates available, influential data can be removed without removing the whole gene; however, when there is a small number of replicates, the entire gene with influential points should be removed from the analysis to preclude bias. More details on *DESeq2*'s features can be found in the study by Love et al.[9] In conclusion, *DESeq2* is recommended by its authors as an improved solution to perform differential analysis because it adopts many competitive features.

Beta binomial. *Overview*. A beta-binomial model is another alternative distribution to accommodate overdispersion.[11,12,42] The beta-binomial distribution has been used in the study by Baggerly et al.[11] to account for both between-library and within-library variations. The authors assumed that the true proportion of gene g within a library i, θ_{gi}, is library-specific and follows a beta distribution: $\theta_{gi} \sim \text{Beta}(\alpha, \beta)$, and that the count Y_{gi} given θ_{gi} follows binomial (m_i, θ_i). Zhou et al.[12] proposed a method, *BBSeq*, which also assumes a beta-binomial distribution and models the proportions of gene g within sample i with a logistic regression. To estimate overdispersion parameters, *BBSeq* either treats the parameter as free and maximizing likelihood directly, or estimates the parameter through modeling the mean-overdispersion relationship.

Modeling. In a beta-binomial model, y_{gi} is converted from the count of gene g in sample i, to proportion, θ_{gi} where $\theta_{gi} = \frac{y_{gi}}{\sum y_{gi}}$. The model is constructed as:

$$\text{logit}(E\boldsymbol{\theta}_{g.})) = \log\left(\frac{E(\boldsymbol{\theta}_{g.})}{1-E(\boldsymbol{\theta}_{g.})}\right) = \boldsymbol{X}\boldsymbol{\beta}_g \qquad (3.3.1)$$

where $\boldsymbol{\beta}_g$ is a vector of the regression coefficients for sample covariates and is the parameter for hypothesis testing; $\boldsymbol{\theta}_{g.}$ is a vector consisting of the proportion of gene g for sample i through N. With the beta-binomial distribution, we are no longer working with a log link but a logit link. $\theta_{gi} \sim$ Beta with $E(\theta_{gi}) = \text{logit}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_g)$ and $\text{var}(\theta_{gi}) = \phi_g E(\theta_{gi})(1-E(\theta_{gi}))$, where ϕ_g is the dispersion parameter. The hypothesis test is constructed as $H_0 : \beta_{gC_1} = \ldots = \beta_{gC_K}$, where β_{gC_K} denotes the estimated coefficient of the indicator variable with 1 for samples in group k and 0 otherwise.

Bayesian and empirical Bayesian. *Overview*. RNA-seq DEA can be modeled in Bayesian framework using various parametric and nonparametric priors. Van de Wiel et al.[13] proposed a Bayesian method, *ShrinkSeq*, which either assumes an informative prior for the overdispersion such as the Dirac–Gaussian prior or estimates one with the empirical Bayesian approach. An empirical Bayesian approach differs from a fully Bayesian approach in that it borrows information from data to elicit priors for overdispersion parameters. For estimating posteriors, Van de Wiel and others[13] adapted the use of integrated nested Laplace approximations, a method that only considers marginal posteriors, but adds a direct maximization of marginal likelihood to allow information sharing from joint posteriors. They further suggested that the use of informative priors for shrinkage, as in *ShrinkSeq*, can ensure stability and accommodate multiplicity correction. They also suggested that shrinkage should be applied not only to overdispersion parameters but also to the regression coefficient parameters. *baySeq*, proposed by Hardcastle and Kelly,[14] constructs the data with tuples grouping genes together based on the study of interest. The distribution of a tuple shares the parameters of some prior distribution so that one can consider many hypotheses for testing beyond two group comparison. The method assumes a negative binomial distribution from the data. *baySeq* first estimates the empirical distribution on the set of parameters for null and alternative models with the quasi-likelihood approach. Then, it estimates the prior probabilities starting from a prior followed by an iterative process updating the priors until convergence. The authors suggested using a log posterior probability ratio of DE for DEA and noted that the posterior probability of DE for each individual model can be conveniently summed up for hypothesis testing.

Modeling. A Bayesian GLM for RNA-seq can be set as:

$$Y_{gi} \overset{d}{=} F_{\mu_{gi},\gamma_g}, \qquad (3.4.1)$$

where γ_g is a vector of parameters not in the regression. The model is in fact flexible in that F can be negative binomial or other distributions. Suppose F follows a negative binomial distribution, then $y_{gi} \sim \text{Poisson}(\mu_{gi})$; μ_{gi} follows a gamma: $\mu_{gi} \sim \text{Gamma}(e^{\eta_{gi}}, \gamma_g)$, where η_{gi} and γ_g are hyperparameters and $\eta_{gi} = \boldsymbol{X}\boldsymbol{\beta}_g = \beta_{g0} + \sum_{p=1}^{P} \beta_{gp} x_{ip}$. x_{ip} is the value of the pth covariate for sample i, such as β_{g1} in a two-group comparison. With $g(\cdot)$ as a link function, $\mu_{gi} = g^{-1}(\eta_{gi})$. The conditional posterior distribution for $\boldsymbol{\beta}$ is proportional with its prior:

$$P(\boldsymbol{\beta} \mid \gamma_g, y_{gi}) \propto P(\boldsymbol{\beta}) \Pi \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})^{y_{gi}}}{1+\exp(\boldsymbol{X}\boldsymbol{\beta})^{y_{gi}+\gamma_g}} \qquad (3.4.2)$$

Each parameter has its respective informative prior and one has to specify priors conditional on the model of interest as well as the prior itself to reach the posterior probability. For testing, a null hypothesis of $\beta_g \leq$ prior under the null is used.

Algorithm Overview 6: Van de Wiel et al.'s[13] *ShrinkSeq*

ShrinkSeq assumes that α is the unknown hyperparameter from a collection of all unknown hyperparameter vectors $\boldsymbol{A}$. It uses a direct maximization of the marginal likelihood method for the estimation of $\boldsymbol{A}$; this method is a modified version of *INLA*.[43] The procedure of finding α is shown below and is said to be analogous to the *EM algorithm*:

1. Initiate $l = 0$ and $\boldsymbol{\alpha}_b^{(0)}$ for $b = 1, \ldots, B$.
2. Use *INLA* to estimate posteriors $\pi_{A^{(l)}}(\theta \mid Y_g)$.
3. Obtain $\boldsymbol{\alpha}_b^{(l+1)}$ for $b = 1, \ldots, B$ with ML′.
4. Iterate from step 2 until convergence.

Notes: let b be the number of informative priors and $\boldsymbol{\alpha}_b^{(l)}$ be the bth element of $A^{(l)}$ at iteration l; let $\pi_{A^{(l)}}$ be the posterior of θ_g condition on data Y_g with $A^{(l)}$ as the current estimate of $\boldsymbol{A}$. ML' is $\boldsymbol{\alpha}^{ML',(l+1)} = \text{argmax}_\alpha \sum_{s=1}^{S} \log(\pi_\alpha(z_{s,A^{(l)}}))$, where this is the prior log-likelihood at $z_{A^{(l)}}$ and s is a large independent sample set from $\overset{\text{EmpBayes}}{\pi}_{A^{(l)}}(\theta)$; ML' has the same mechanism as the maximum likelihood.

Dirac–Gaussian and Gaussian–Dirac–Gaussian mixture priors:

$$\pi(\beta) = p_0\delta_0 + (1-p_0)N(\beta;0,\tau^2), \quad (3.4.a)$$

$$\pi(\beta) = p_{-1}N(\beta;\mu_{-1},\tau_{-1}^2) + p_0\delta_0 + p_1N(\beta;\mu_1,\tau_1^2), \quad (3.4.b)$$

The subscripts of p, ie, –1, 0, and 1, indicate the locations. For example, Dirac mass on 0 is denoted as δ_0. Considering the p as probability where p_{-1}, p_0, and p_1 sum up to 1, then $p_0 = 1 - p_{-1} - p_1$. $\mu_{-1} < 0$, $\mu_1 > 0$. Priors with positive mass on zero were intentionally selected because it reflects the non-DE condition. For more details on priors, please refer to the study by Van de Wiel et al.[13]

Algorithm Overview 7: Hardcastle and Kelly's[14] *baySeq*

The tuple system in *baySeq* is as follows. Let a model be denoted as M. E refers to a set of models described by the data, $\{E_1 \ldots E_l\}$. κ represents the set of parameters for each model, M, ie, $\{\theta_1 \ldots \theta_l\}$. Let q be the index of each underlying distribution for model 1, …,l. An example would be that samples in groups 1, 2, and 3 (C_1, C_2, C_3) are grouped together in a way that groups 1 and 2 are equivalently distributed and group 3 stands alone: $M = \{A_{i\in C_1}, A_{i\in C_2}\}$, $\{A_{i\in C_3}\}$ where A is the sample. D_t is the data in tuple t : $\{\{y_{1t} \ldots y_{it} \ldots y_{n_t t}\}, \{m_1 \ldots m_i \ldots m_{n_t}\}\}$, which is the count in tuple t for sample i, m_i is the library size. The posterior probability of model given data is:

$$P(M \mid D_t) = \frac{P(D_t \mid M)P(M)}{P(D_t)} \quad (3.4.c)$$

$$\text{where} \quad P(D_t \mid M) = \int P(D_t \mid \kappa, M)P(\kappa \mid M)d\kappa \quad (3.4.d)$$

Suppose that a sample A_i is in the set E_q where the count of this sample at a particular tuple t is y_{it}, which follows a negative binomial(μ_{it}, φ_q) ($\theta_q = (\lambda_q, \varphi_q)$). The mean count μ_{it} is a product of the library size scaling factor, m_i, and the proportion of reads in set E_q, λ_q. We have:

$$P(D_t \mid \kappa, M) = P(y_{it} \mid m_i, \theta_q) = \frac{\Gamma(y_{it} + \phi_q^{-1})}{\Gamma(\phi_q^{-1})\Gamma(y_{it}+1)} \left(\frac{1}{1+\mu_{it}\phi_q}\right)^{\phi_q^{-1}} \left(\frac{\mu_{it}}{\phi_q^{-1}+\mu_{it}}\right)^{y_{it}} \quad (3.4.e)$$

baySeq first estimates the empirical distribution on the set of parameters for null and alternative models through sampling from a negative binomial distribution and a quasi-likelihood approach.[38] Then, it estimates the prior probabilities starting from a prior followed by an iterative process updating the priors until convergence. For detailed steps, please refer Hardcastle and Kelly.[14] Hypothesis testing can be easily denoted with the tuple system, for instance a two-group case,

$$H_0(\text{non-DE}): \{A_{i\in C_1}, A_{i\in C_2}\}$$
$$H_a(\text{DE}): \{A_{i\in C_1}\} \text{ and } \{A_{i\in C_2}\}$$

Nonparametric. *Overview*. In this section, we discuss two nonparametric methods for RNA-seq DEA by Li and Tibshirani[15] and Tarazona et al.[16] In *SAMseq*, Li and Tibshirani[15] calculated a modified two-sample Wilcoxon statistic using the ranked counts for two-group comparison.[44] The authors proposed two resampling strategies for producing equal sequencing depths of the samples: downsampling and Poisson sampling, and also suggested that ties can be broken by inserting a small random number in resampling. *NOISeq* by Tarazona et al.[16] first used pseudo-counts corrected by the library size $m_{k(i)}$ under two conditions (K=2) to calculate log-ratio (M) and absolute value of difference (D). Then, a test statistic is derived from M and D with a null hypothesis of no differential expression; in other words, M and D are no different than random variables either estimated from the real or simulated data.

Modeling. The two nonparametric methods discussed here are explained separately in the test boxes, as they each has a unique model setup.

Algorithm Overview 8: Li and Tibshirani's[15] *SAMseq*

To use *SAMseq*, one ranks the counts of gene g across samples and denotes the ordered counts as $y'_{g1} \ldots y'_{gN}$. If needed, resampling strategy may be used to fulfill the requirement of equal sequencing depths of samples in Wilcoxon test.

In the case of a sufficient minimal sequencing depth, the authors proposed a downsampling strategy where one first identifies the smallest sequencing depth, denoted as $m_{\min}$, where $m_{\min} = \min(m_1, \ldots, m_N)$ and keeps this list

of counts while resampling lists of counts for all other samples with the sequencing depth, $m_{\min}$. Every count is randomly sampled with a success probability of $m_{\min}/m_i$ and failure probability of its complement, ie, the resampled count is

$$y'_{ij} \sim \text{binomial}\left(y_{ij}, \frac{m_{\min}}{m_i}\right). \quad (3.5.a)$$

In the case of an insufficient minimal sequencing depth, Li and Tibshirani[15] introduced Poisson sampling strategy, wherein they employed the geometric mean of the sequencing depths for all samples:

$$y'_{ij} \sim \text{Poisson}\left(\frac{\bar{m}}{m_i} N_{ij}\right), \quad (3.5.b)$$

where y'_{ij} is resampled data and $\bar{m} = (\prod_{i=1}^{N} m_i)^{1/N}$. Small random numbers are introduced into the resampling process to break ties, as well as multiple resampling to ensure stability. Poisson sampling is generally preferred based on the simulation.[15] In cases where m_i is unknown, one could use normalization methods to estimate. Differential expression of gene g is identified based on a comparison of the ranks of gene g between the two sample groups.

Algorithm Overview 9: Tarazona et al.'s[16] *NOISeq*

In *NOISeq*, for each $y_{gk(i)}$, the count of gene g in sample i from group k, the correction method for library size, $m_{k(i)}$, is the sum of counts over all genes for the ith sample replicate in condition k. Let $m_{k(i)}$ be simplified as m_i. One would work with pseudocounts (after normalization) formulated as: $\tilde{y}_{gk(i)} = y_{gk(i)} \times 10^6 / m_i$.

With the pseudocounts, the log ratio (L) and the absolute value of difference (D) are calculated. $\tilde{y}_{gk}$ is summarized over ith samples, a.k.a. $\tilde{y}_{gk} = \sum_{i \in C_k} \tilde{y}_{gi}$. $L_g = \log_2\left(\frac{\tilde{y}_{gC_1}}{\tilde{y}_{gC_2}}\right)$ and $D_g = |\tilde{y}_{gC_1} - \tilde{y}_{gC_2}|$, where C_1 and C_2 denote group 1 and 2, respectively. Zero counts are replaced by 0.5 or by mid(0, normalized minimum expression) when calculating L_g. Samples with only zeros are dropped.

Null hypothesis: L and D values are no different than noise if no DE. Probability distribution for random variables L^* and D^* are either estimated from real data or simulated data and are used for the noises. One then obtains the probability of DE as:

$$\begin{aligned} &P(DE_g = 1 | \tilde{y}_{gC_1} - \tilde{y}_{gC_2}) \\ &= P(DE_g = 1 | L_g = l_g, D_g = d_g) \\ &= P(|L^*| < |l_g|, D^* < d_g) \end{aligned} \quad (3.5.c)$$

DE_g equals 1 when gene g is DE. Note that log ratio is in absolute term because either direction indicates DE. See the study by Tarazona et al.[16], for more details.

Statistical Testing

After performing parameter estimation for a statistical model, significance of differential expression can be assessed comparing the expression of gene g among K groups. Assume that $\lambda_{gk(i)}$ is the expression level of gene g in sample i belonging to sample group k. ϕ_g is the dispersion parameter. DE tests are proposed below for the null hypothesis (H_0):

$$\lambda_{g1} = \ldots = \lambda_{g\kappa}.$$

In parametric regime, one can employ classic log-likelihood ratio test.

$$LR_g = \frac{2(l_g(\hat{\lambda}, Y_g) - l_g(\hat{\lambda}^0, Y_g))}{\hat{\phi}_g} \sim F_{K-1,N-K} \quad (4.1)$$

In absence of overdispersion,

$$LR_g = 2(l_g(\hat{\lambda}, Y_g) - l_g(\hat{\lambda}^0, Y_g)) \sim \chi^2(K-1) \quad (4.2)$$

where l_g denotes the log-likelihood function for the gth gene; $l_g(\hat{\lambda}, Y_g)$ and $l_g(\hat{\lambda}^0, Y_g)$ denote the MLE of biological and experimental effects under the full model and null model, respectively.

An exact test for negative binomial, analogous to the Fisher's exact test, is used by methods, such as *edgeR* and *DESeq*. By conditioning on the total sum, one can calculate the probability of observing counts as extreme or more extreme than what is really obtained, resulting in an exact P-value. Note that a sum of gene counts from all replicates in each group that is either too large or too small indicates a differential expression, so a two-sided test is used.

A score statistic is used by *PoissonSeq*, which tests for the significance of the association of gene g with expression of groups. In the context of gene count with unknown dispersion parameters, a score test is as follows:

$$S_g = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{w_g (y_{gi} - \hat{\mu}_{gi})^2}{\phi_g v(\hat{\mu}_{gi})} \sim F_{K-1,N-K} \quad (4.3)$$

Table 2. RNA-seq count data DEA statistical methods and software.

MODEL	SOFTWARE	REFERENCES	DATA TYPE	TESTING STRATEGY	NOTES	LIMITATIONS	DATA USED
Poisson	DEGseq	Wang et al.[4]	RNA-seq data	Fisher's exact test Likelihood ratio test	Support raw read counts or normalized gene expression values, identify DE of exons or transcripts	Ignore biological variation	Marioni RNA-seq data
	Myrna	Langmead et al.[5]	RNA-seq data	Likelihood ratio test Parallelized permutation test	Handle dataset with over 1 billion rows, computationally efficient	Ignore biological variation, signal loss due to junction or repetitive reads, inconvenient cloud data transfer	HapMap expression data
	PoissonSeq	Li et al.[6]	RNA-seq data Tag-seq data	Score test	Accommodate multiple covariate types, computationally efficient	Transformation power depends only on gene expression, libraries are totally exchangeable	Marioni RNA-seq data Tag-seq data
Negative binomial	edgeR	Robinson et al.[7]	SAGE data RNA-seq data	Exact test Likelihood ratio test	Separate biological from technical variations	Limited to pairwise comparison	SAGE data Fly RNA-seq data
	DESeq	Anders and Huber[8]	Tag-seq data RNA-seq data ChIP-seq data	Exact test Likelihood ratio test	Extend edgeR by allowing more general, data-driven relationship of mean and variance	Limited to pairwise comparison	Neural stem cell Tag-seq data Yeast RNA-seq data HapMap ChIP-seq data Fly RNA-seq data
	DESeq2	Love et al.[9]	Tag-seq data RNA-seq data ChIP-seq data	Wald test Likelihood ratio test	Improve upon DESeq for better gene ranking, allow hypothesis tests above and below threshold	Limited to pairwise comparison	Fly RNA-seq data Mouse straiturn RNA-seq data
	NBPSeq	Di et al.[10]	RNA-seq data	Adapted exact test	Introduce an additional parameter to allow the dispersion to depend on the mean	Assume all library sizes are equal	Arabidopsis RNA-seq data
Beta binomial	BBSeq	Zhou et al.[12]	RNA-seq data	Wald test Likelihood ratio test	Handle outlier detection automatically	Sensitive to outliers of shrinkage or penalization methods	HapMap RNA-seq data
Bayesian and Empirical Bayesian	ShrinkSeq	Van de Wiel et al.[13]	RNA-seq data CAGE data	Evaluating posterior probability for inference	Provide joint shrink multiple parameters, allow for random effects, address multiplicity problems	Computationally intensive but allow parallelization	HapMap RNA-seq data CAGE data
	baySeq	Hardcastle and Kelly[14]	Small RNAs data	Evaluating posterior probability for inference	Involve multiple comparison, accommodate different sample size	Computationally intensive but allow parallelization	Trans-acting small RNAs
Non-parametric	SAMseq	Li and Tibshirani[15]	RNA-seq data Tag-seq data miRNA-seq data	Wilcoxon test	Robust to outliers, remove Experimental effect, simplify test for feature effect, accommodate quantitative, survival and multiple group comparison	Overestimate FDR in some cases, relative low power for data with small sample size	Marioni RNA-seq data t'Hoen Tag-seq data Witten miRNA-seq data
	NOIseq	Tarazona et al.[16]	RNA-seq data	Wilcoxon test	Robust and maintain a high true-positive rate	Not easy to identify true differential expression at a low count range, limited to pairwise comparison	Marioni RNA-seq data

where w_g is a known weight, $\hat{\mu}_{gi}$ is estimated by MLE under the null hypothesis, and $v(\hat{\mu}_{gi})$ is the variance function of μ_{gi}.

Wilcoxon statistic is a rank-transformed version of *t*-statistics, used by the nonparametric method, *SAMseq*:

$$W_g = \sum_{i \in C_k} r_{gi} - r_0, \tag{4.4}$$

where r_{gi} is the rank of y_{gi} across samples and $r_0 = (\sum I_{(i \in C_k)})(N+1)/2$ (r_0 is used to make $E(W_g) = 0$). $W_g > 0$ identifies that gene g is overly expressed in group k.

Under a Bayesian or empirical Bayesian framework, methods like *baySeq* use posterior likelihood of the DE model per gene to identify differential expression:

$$P(M_{H_0} | Y_g) = \frac{P(Y_g | M_{H_0}) P(M_{H_0})}{P(Y_g)}, \tag{4.5}$$

where *M* denotes a model. Posterior probability of DE to non-DE ratio is often used.

The choice of a testing strategy is a decision that often depends on the chosen method and other factors such as sample size. With a small sample size, the large-sample approximations based on the Wald test, score test, and likelihood ratio test are questionable and an exact test is usually preferred.[36] We summarize testing strategies that are plausible for each method in Table 2.

Finally, almost all the methods we mentioned in this paper use standard approaches for multiple hypothesis correction to control false discoveries.[45,46] *PoissonSeq* is an exception that builds its own estimation of false discovery rate (FDR) from a permutation test. Permutation test calculates a score test per gene, S_g, for H_{0g} vs H_{ag}, each time when the outcome is permuted. For *B* permutations, the same procedure is applied to calculate null statistics S_g^{0b} for $b = 1 \cdots B$. The permutation *P*-value is:

$$p_g = \sum_{b=1}^{B} \sum_{i=1}^{N} \frac{I\{S_g^{0b} > S_g\} + 1}{N \times B + 1} \tag{4.6}$$

For Bayesian methods, since posterior probabilities are computed, Bayesian FDR or local FDR are conveniently used. Local false discovery rate ($lFDR_g$) is simply the posterior probability π_{0g}:

$$lFDR_g = P(M_{H_0} | Y_g) = P_0 / (P_0 + P_\alpha), \tag{4.7}$$

where $P_0 = \int_{-\infty}^{\triangle} P(Y_g | \lambda_g = \lambda) \pi(\lambda) d\lambda$, and $P_1 = \int_{\triangle}^{\infty} P(Y_g | \lambda_g = \lambda) \pi(\lambda) d\lambda$, and $\triangle$ denotes prior. Bayesian false discovery rate (BFDR) is calculated as:

$$BFDR(t) = \frac{\sum_{g=1}^{G} lFDR_g \times I\{\pi_{0g} < t\}}{\sum_{g=1}^{G} I\{\pi_{0g} < t\}}. \tag{4.8}$$

Note that $I\{\pi_{0g} < t\} = I\{\pi_{1g} \geq t\}$ for small t of interest.

Conclusion

RNA-seq data analysis is a relatively new and rapidly growing research area. The statistical model used for sequencing data has been evolving. The first proposed Poisson distribution has become obsolete because it fails to accommodate commonly-observed overdispersion in RNA-seq data. In a parametric framework, the negative binomial distribution is the most common assumption for modeling the marginal distribution due to the technical and biological variations.[8,9,33,36] Other available methods that account for overdispersions include the generalized Poisson distribution,[35] negative binomial power distribution,[10] and beta-binomial distribution,[11,12] as well as nonparametric models[15,16] and Bayesian methods.[13,14] Table 2 summarizes all the reviewed methods in this paper.

For readers who are interested in the performance evaluation and method comparison of the available methods, they can refer to the original paper as well as the body of literature on this issue. For instance, in the study by Seyednasrollah et al.[22], *DESeq* has been recommended as one of the most robust methods and caution is advised when dealing with a small number of replicates regardless of which method is being used. Similarly, Soneson and Delorenzi[21] also advise caution when interpreting results drawn from a small number of replicates and show that *SAMseq* surpasses many other reviewed methods. In the study by Rapaport et al.[23], *DESeq*, *edgeR*, and *baySeq*, which all assume a negative binomial model, have better specificity, sensitivity, and control of false positive errors than other nonnegative binomial models. As the technology continues to improve and the empirical data accumulate, more compelling statistical modeling for RNA-seq data can be expected.

Author Contributions

Conceived and designed the experiments: HCH, YN, LXQ. Reviewed the literature: HCH, YN, LXQ. Wrote the first draft of the manuscript: HCH, YN. Contributed to the writing of the manuscript: HCH, YN, LXQ. Agree with manuscript results and conclusions: HCH, YN, LXQ. Jointly developed the structure and arguments for the paper: HCH, YN, LXQ. Made critical revisions and approved final version: HCH, YN, LXQ. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63.
2. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*. 2011;9:34.

3. Dillies MA, Rau A, Aubert J, et al; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*. 2012;14(6):671–83.
4. 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;26:136–8.
5. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*. 2010;11:R83.
6. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*. 2012;13:523–38.
7. Robinson MD, McCathy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.
8. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R25.
9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
10. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial models for assessing differential gene expression from RNA-seq. *Stat Appl Genet Mol Biol*. 2011;10:1.
11. Baggerly KA, Deng L, Morris JS, Aldaz CM. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*. 2003;19(12):1477–83.
12. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*. 2011;27:2672–8.
13. Van de Wiel MA, Leday GG, Pardo L, Rue H, Van de Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*. 2013;14:113–28.
14. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*. 2010;11:422.
15. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. *Stat Methods Med Res*. 2011;22(5):519–36.
16. Tarazona S, Garcia-Alcalde F, Ferrer A, Dopazo J, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res*. 2011;21:2213–23.
17. 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;18:1509–17.
18. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
19. Oshlack A, Robinson MD, Young MD. From RNA-Seq reads to differential expression results. *Genome Biol*. 2010;11:220.
20. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot*. 2012;99:248–56.
21. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013;14:91.
22. Seyednasrollah F, Laiho A, Elo L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*. 2015;16(1):59–70.
23. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013;14:R95.
24. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11:733–9.
25. Chen Y, McCarthy D, Robinson M, Smyth GK. edgeR: differential expression analysis of digital gene expression data. User's Guide; http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf 2015.
26. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621–8.
27. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11:R25.
28. Anders S, McCarthy D, Chen Y, et al. Count-based differential expression analysis of RNA sequencing data using R and bioconductor. *Nat Protoc*. 2013;8:1765–86.
29. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*. 2011;12:480.
30. Benjamini Y, Speed T. Estimation and correction for GC-content bias in high throughput sequencing. *Nucleic Acids Res*. 2011;40(10):e72.
31. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. 2012;13(2):204–16.
32. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19:185–93.
33. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23:2881–7.
34. Auer PL, Deroge RW. A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol*. 2011;10:26.
35. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res*. 2010;38:e170.
36. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion with applications to SAGE data. *Biostatistics*. 2008;9:321–32.
37. Smyth GK. Pearson's goodness of fit statistic as a score test statistic. In: Goldstein DR, ed. *Science and Statistics: A Festschrift for Terry Speed*. Hayward, CA: Institute of Mathematical Statistics; 2003:115–26. [*IMS Lecture Notes Monograph Series* 40].
38. Nelder JA, Lee Y. Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *J Roy Stat Soc B*. 1992;54:273–84.
39. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. *J Roy Statist Soc B*. 1987;49(1):1–39.
40. Dimont E, Shi J, Kirchner R, Hide W. edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test. *Bioinformatics*. 2015;31(15):2589–90.
41. Reid N. The roles of conditioning on inference. *Stat Sci*. 1995;10(2):138–57.
42. Zhang L, Zhou W, Velculesu VE, et al. Gene expression profiles in normal and cancer cells. *Science*. 1997;276(5316):1268–72.
43. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J Roy Stat Soc B*. 2009;71:319–92.
44. Wilcoxon F. Individual comparisons by ranking methods. *Biometric Bulletin*. 1945;1(6):80–3.
45. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995; Series B 57(1):289–300. MR 1325392.
46. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat*. 2002;31(6):2013–35.