Proposal

May 22, 2019

Contents

Abstract

1 Introduction

Ribonucleic acid (RNA) is an essential biomolecule for all forms of known life on Earth. RNA is assembled as a chain of nucleotides that are made up of a base, a sugar, and at least one phosphate group. The genetic information of RNA is stored in the bases guanine (G), uracil (U), adenine (A), and cytosine (C). There are many different types of RNA associated with different functions (Clancy, 2018). For example, both messenger RNA (mRNA) and translation RNA (tRNA) are directly involved in the protein synthesis, one of the most fundamental biological processes by which individual cells build their specific proteins.

RNA-sequencing (RNA-Seq) makes use of next generation sequencing (NGS) technology to identify and quantify different types of RNA at a massive scale (Ozsolak and Milos, 2010; Chu and Corey, 2012). This technology provides genomic information about samples, helps us understand how populations differ at the genomic level, makes personalized genomics possible, and plays an important role in research areas such as cancer and agriculture (Ozsolak and Milos, 2010; Anderson and Habiger, 2012).

NGS RNA-Seq data is an $M \times N$ matrix of counts $\boldsymbol{Y}$. Here Y_{mn} corresponds to data for the m^{th} feature in the n^{th} sample. Typically, M is normally tens of thousands while N is small ranges from 2 to 10. Data are typically from 2 or more populations or treatment groups or there may be a quantitative trait measurement, or covariate x_n, from the n^{th} sample. The objective of the data analysis is to identify the features that are differentially expressed across different populations or treatment groups or identify the features that are associated with the trait variable of interest. Our focus is to study which features are associated with the trait.

The standard approach is to test the null hypothesis that the m^{th} feature is not associated with the trait for $m = 1, 2, \ldots, M$. A Poisson log linear regression model can be used to get a z-score and p-value from each null hypothesis. Then a multiple testing procedure such as the local false discovery rate ($lFDR$) procedure or the Benjamini and Hochberg (BH) procedure can be applied to the collection of of z-scores and p-values to determine which hypotheses are rejected.

There are three known statistical challenges in this standard approach:

- Overdispersion occurs. That is the variance is often larger than the mean. The Poisson model can underestimate the variance since it assumes that the variance and mean are equal. This can inflate the type I error rate (Kvam, Liu, and Si, 2012; McCarthy, Chen, and Smyth, 2012; Anders and Huber, 2010).

- The total number of reads in a sample is heterogeneous across samples because each sample is sequenced to different depths. This leads to different expected number of counts for m^{th} feature in n^{th} sample. This effect can be confounded with the effect caused by differential expression (J. Z. Li and Tibshirani, 2013).

- The total number of reads for a feature is heterogeneous across features. If we ignore heterogeneity, then the features with large total number of reads, rather than the ones that are most strongly associated with the trait, will be identified (Habiger, Watts, and Anderson, 2017).

A lot of research has addressed one or two of the above issues. We propose methods that address all three statistical challenges simultaneously. We expect that the new methods will identify more features that are strongly associated with the trait than existing methods without sacrificing control of the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

Section 2 provides a literature review. NGS technology is introduced in Section 2.1 and standard RNA-Seq data analysis approaches are in Section 2.2.

A real data application is in Section 2.3. A detailed explanation and discussion of the three statistical challenges and an extensive literature review on the methods for them are in Section 2.4. Section 3 proposes the research plan. All three statistical challenges are discussed in Section 3.1. Analytical assessment methods are proposed in Section 3.2. Simulation studies are proposed in Section 3.3.

2 Literature Review

In this section we provide a brief review of NGS technology. An introduction to the analysis of NGS data with a real example follows.

2.1 Introduction to Next Generation Sequencing

This subsection briefly introduces RNA-Seq and the basic procedure of NGS technology. The basic NGS procedure is detailed below. See Figure 1 for a depiction.

1. A relevant biological sample of tissues or cells is collected from each experiment or observational unit. An RNA sample is extracted from each of the biological samples. Then they are converted to a sample of DNA or complementary DNA (cDNA) fragments. The fragments are ligated by specialized adapters either on one or both ends. The resulting sample of fragments is often referred to as a library, which is ready for sequencing. This step is called library preparation (Z. Wang, Gerstein, and Snyder, 2009).

2. The library is loaded into one of the eight lanes on a hollow glass slide called the flow cell based on library adapters (Mitra et al., 2014). It is amplified and sequenced by a high-throughput sequencing platform. Technical replicates can be obtained by sequencing the same library more than once in the same flow cell (Nettleton, 2014).

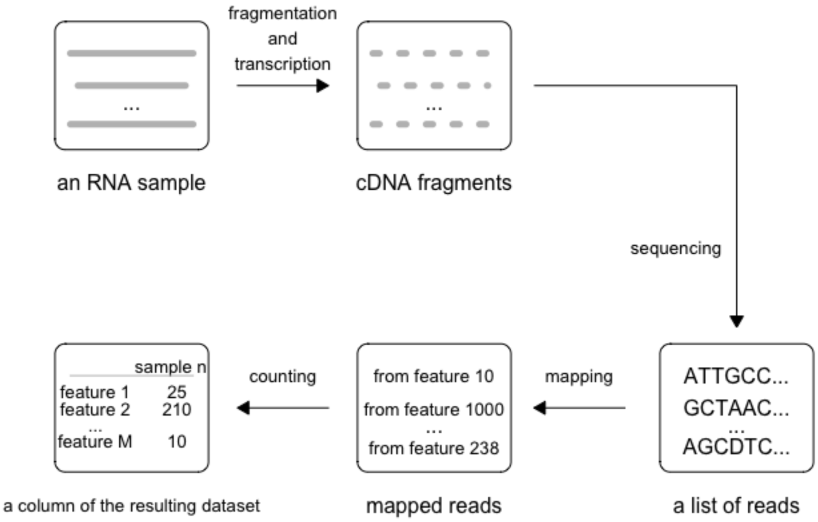

Figure 1: Next generation sequencing technology procedure

3. Either reads (the sequences of bases from DNA strands, defined in Goodwin, Mcpherson, and Mccombie, 2016) are mapped to the reference genome or the de novo assembly method is used to reconstruct the original sequence (Mitra et al., 2014). Millions of reads are generated from each library.

4. The reads are mapped and aggregated to features of interest (such as genes, transcripts, or operational taxonomic units). The number of reads assigned to each feature is called counts, which gives a measure of abundance for the feature in the library studied (Finotello and Di Camillo, 2014). The end results are the features of interest and the counts of the features. The total number of counts for each library is called library size (Lorenz et al., 2014).

The error rates associated with NGS technologies are $\sim 0.1-15\%$ (Goodwin, Mcpherson, and Mccombie, 2016).

2.2 Introduction to RNA-Seq data analysis

In this section we introduce some basic definitions, criteria for simultaneous hypotheses testing and related procedures in RNA-Seq data analysis.

2.2.1 Objective

This subsection highlights our statistical analysis objective by reviewing some standard criteria used in single and multiple hypothesis testing.

Recall that in a typical RNA-Seq data study the goal is to identify features that are associated with a quantitative trait of interest or a treatment. To accomplish this objective, the null hypothesis that a feature is not associated with the trait or a treatment is tested against the alternative hypothesis that an association exists. If the null hypothesis is rejected when it is true, a Type I error is committed. A Type I error is also called a false positive or a false discovery. When a feature is claimed not to be associated with the quantitative trait when it actually is, a Type II error is committed. A Type II error is called a false negative.

For a collection of M tests, let M_0 denote the number of null hypotheses that are true. The number of Type I errors or false positives is denoted as V and the number of Type II errors or false negatives is denoted as T. The number of true null hypotheses that are not rejected is denoted as U, and the number of true non-null hypotheses that are rejected is denoted as S. Let R be the total number of rejected null hypotheses. Table 1 from Benjamini and Hochberg (1995) has summarized the notation. The objective is to maximize the expected number of true positives S while safeguarding against some global Type I error rate, typically depending on V and/or R.

	Non-rejected null	Rejected null	Total
True null	U	V	M_0
True non-null	T	S	$M - M_0$
Total	$M - R$	R	M

Table 1: R of the M null hypotheses are rejected; V of them are false rejections, while S of them are correct.

In classical hypothesis testing of a single null hypothesis, the null hypothesis is rejected when the p-value is less than or equal to α, so that the probability of committing a type I error is controlled at level α. However, when there are multiple tests, the expected number of false positives αM can be very large. For example, 5 false discoveries are expected when 100 true null hypotheses are tested at level $\alpha = 0.05$.

The family-wise error rate ($FWER$) is defined as $FWER = P(V \geq 1)$. One standard procedure to control the $FWER$ is Bonferroni's procedure, which rejects the m^{th} hypothesis if $p_m \leq \frac{\alpha}{M}$. However, the $FWER$ was originally developed for a small number of hypotheses tests, say $M \leq 20$ (Efron and Hastie, 2016). When M increases to thousands or even millions, $FWER$ methods are generally too conservative.

2.2.2 False discovery rate methods

This subsection introduces the false discovery rate (FDR) and two common procedures to control it. Let $Q = \frac{V}{R}$ be the proportion of false rejections among all the rejected null hypotheses. The FDR is defined as the expectation of Q. It has been shown that $FDR = E(Q) \leq P(V \geq 1) = FWER$ (Benjamini and Hochberg, 1995) and hence FDR methods are generally more liberal.

The Benjamini and Hochberg (BH) procedure (Benjamini and Hochberg, 1995) is a commonly used method for FDR control. It is implemented as follows.

1. Order the P-values $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(M)}$.
2. Find $j = \max\{m : P_{(m)} \leq \frac{m}{M}\alpha\}$.
3. Reject all $H_{(m)}$, where $m = 1, 2, \cdots, j$ and $H_{(m)}$ is the null hypothesis corresponding to $P_{(m)}$.
4. If $P_{(m)} > \frac{m}{M}\alpha$ for all m, no null hypotheses are rejected.

Another procedure is based on a mixture model (Cai and Sun, 2009; Efron, 2008). Assume the proportion of true null hypotheses is p_0, and the proportion of false null hypotheses is p_1. Assume the Z-values corresponding to the null hypothesis follow a standard normal density $f_0(z)$, and those corresponding to nonnull hypotheses have density $f_1(z)$. The mixture density is $f(z) = p_0 f_0(z) + p_1 f_1(z)$, and the local false discovery rate ($lFDR$) is defined in Efron (2008)

$$lFDR(z) = \frac{p_0 f_0(z)}{f(z)}, \tag{1}$$

and the $lFDR$ statistic for H_m is $lFDR_m = lFDR(Z_m)$. The $lFDR$ procedure is (Sun and Cai, 2007):

1. Order the $lFDR$ statistics $lFDR_{(1)} \leq lFDR_{(2)} \leq \cdots \leq lFDR_{(M)}$.
2. Find $j = \max\{m : \sum_{i=1}^{m} lFDR_{(i)} \leq m\alpha\}$.
3. Reject all $H_{(m)}$, where $m = 1, 2, \cdots, j$.
4. If $\sum_{i=1}^{m} lFDR_{(i)} > m\alpha$ for all m, no null hypotheses are rejected.

The above procedure is sometimes called the oracle $lFDR$ procedure because f and p_0 are unknown in practice. Hence, only an oracle who knows f and p_0 can implement it. However, f and p_0 can be estimated and an adaptive oracle procedure can be implemented. The adaptive oracle procedure is also sometimes called an empirical Bayes procedure because the $lFDR$ can be viewed as a posterior probability, hence the plugging in estimates of p_0 and f can be viewed as plugging in estimates of prior parameters. Efron (2008) has shown that the null distribution f_0 can also be estimated in large scale testing situations, and performs better than theoretical null in some scenarios.

2.3 A real example

In this section the BH procedure and adaptive $lFDR$ procedure are applied to an agriculture RNA-Seq dataset.

It is known that rhizobacteria, which are bacteria living near the roots of plants, influence plant productivity (Anderson and Habiger, 2012). In order to identify the rhizobacteria species that are associated with plant productivity, 16S rRNA genes were sequenced from extracted genomic DNA (Tringe et al., 2005) by Illumina MiSeq series. Similar 16S rRNA sequences were grouped into operational taxonomic units (OTUs) (Ye, 2011; Du and Fang, 2014). The number of reads for 1592 OTUs across 9 soil samples were recorded. See Table 2. The shoot biomass from each sample was also recorded as a measure of plant productivity.

The specific objective is to identify the OTUs that are associated with plant productivity. To be more precise, let us introduce some notation. Let y_{mn} denote the number of reads mapped to the m^{th} OTU in the n^{th} sample, where $m = 1, 2, ..., M = 1592$ and $n = 1, 2, ..., N = 9$. Let $y_{m.} = \sum_{n=1}^{N} y_{mn}$ and $y_{.n} = \sum_{m=1}^{M} y_{mn}$ denote the total number of reads for each OTU and each sample, respectively. The vector of reads for the m^{th} OTU is denoted by $\boldsymbol{y_m} = (y_{m1}, ..., y_{mN})^T$. The corresponding random variable for y_{mn} is denoted by Y_{mn} and the corresponding random vector for $\boldsymbol{y_m}$ is denoted by $\boldsymbol{Y_m}$.

Denote shoot biomass for each sample as $\boldsymbol{x} = (x_1, \cdots, x_9) = (7.58, \cdots, 10.53)$. Hence, the objective is to determine if the distribution of $\boldsymbol{Y_m}$ depends on $\boldsymbol{x}$. That is, we want to test the null hypotheses that $\boldsymbol{Y_m}$ doesn't depend on $\boldsymbol{x}$ for $m = 1, 2, \cdots, M$.

OTU m	y_{m1}	y_{m2}	$\cdots$	y_{m9}	Total ($y_{m.}$)
1	4230	3563	$\cdots$	1954	33243
2	3523	3222	$\cdots$	1559	30809
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1592	0	2	$\cdots$	0	10
Total ($y_{.n}$)	81839	67861	$\cdots$	32073	638049

Table 2: Depiction of the OTU data

Assume that Y_{mn} are independent variables that follow Poisson distributions with mean μ_{mn}, and $log(\mu_{mn}) = \alpha_m + \beta_m x_n$. Observe if $\beta_m = 0$, then $\mu_{mn} = e^{\alpha_m}$ for $n = 1, 2, \cdots, 9$, which implies the distribution of $\boldsymbol{Y_m}$ does not depend on $\boldsymbol{x}$. Hence, the null hypothesis for the m^{th} OTU is $H_m : \mu_{m1} = \cdots = \mu_{m9}$ or $H_m : \beta_m = 0$.

A variety of z-scores could be considered. For the sake of illustration, we consider a Wald z-score given by

$$z_m = \frac{\hat{\beta}_m}{(\mathbf{x}^T \hat{\mathbf{W}} \mathbf{x})^{-1}_{mm}}, \tag{2}$$

where $\hat{\mathbf{W}}$ is the diagonal matrix of working weights from standard GLM theory, and $\hat{\beta}_m$ is the maximum likelihood estimate of β_m (McCullagh and Nelder, 1989). The p-value for m^{th} hypothesis is computed as

$$p_m = P(|Z_m| \geq |z_m|) = 2[1 - \Phi(|z_m|)], \tag{3}$$

where $\Phi(.)$ is the cumulative distribution function for standard normal distribution. The z-scores and p-values can be easily computed using the glm function in R.

The two procedures outlined in the previous section can now be applied here. Both procedures are applied at FDR level $\alpha = 0.05$. The BH procedure rejects 328 hypotheses and hence 328 OTUs are discovered. See figure 2 for a depiction of the distribution of p-values. The empirical Bayes procedure rejects 101 null hypotheses and 101 OTUs are discovered. The default setting of Efron's package locfdr has been used when estimating p_0 and f. See figure 3 for a depiction of the distribution of Z-values. However, as we will see in the next section, this standard approach has several drawbacks.

2.4 Statistical challenges in RNA-seq data analysis

There are three statistical challenges that were ignored in the standard approach: overdispersion, heterogeneous library sizes, and heterogeneous total feature counts.

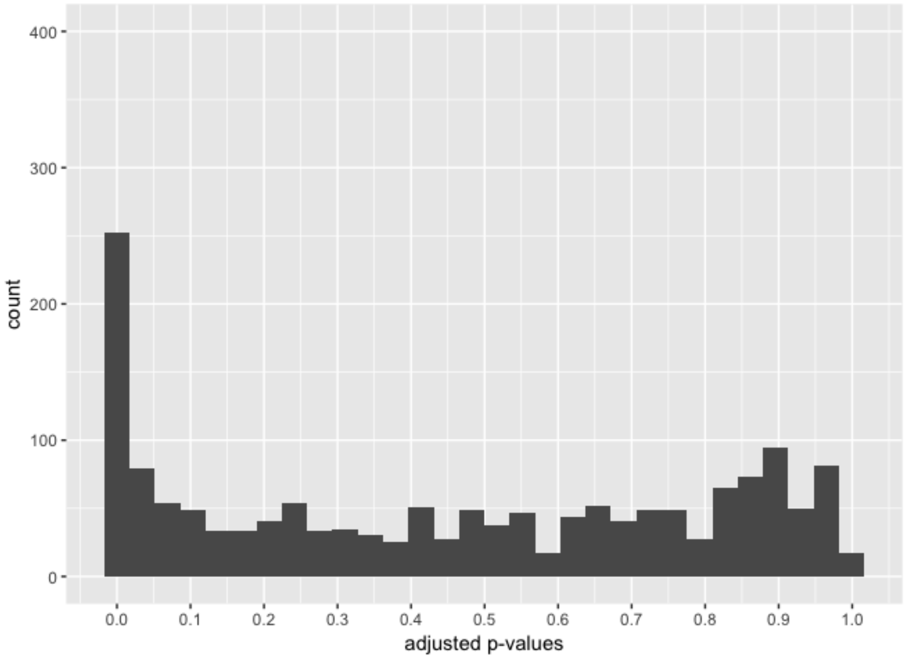

Figure 2: Histogram of adjusted P-values from Wald tests

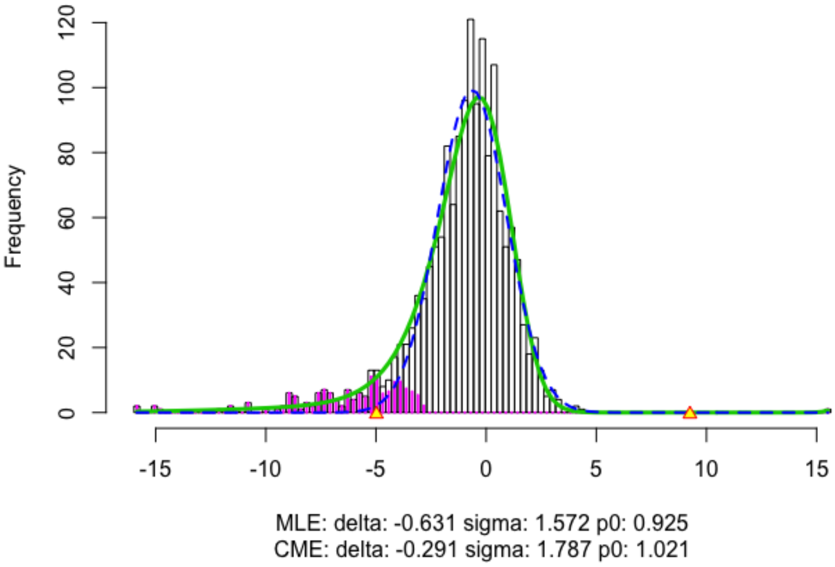

Figure 3: Histogram of z-values with f_0 and $\hat{f}$ superimposed

2.4.1 Statistical Challenge 1 – overdispersion

When we sequence the same sample multiple times, the variation is mainly due to the measurement technology and is called technical variation (Nettleton, 2014). When different experimental units are treated alike or observational units are observed under the same environment, the variation between the units is called biological variation. The total variation, which includes both types of the aforementioned variation, in RNA-Seq data resulting in overdispersion (Robinson and Smyth, 2007; Robinson, McCarthy, and Smyth, 2010; Auer and Doerge, 2011; Di et al., 2011; Kvam, Liu, and Si, 2012; Lund et al., 2012; Chen, Lun, and Smyth, 2014). Several common contributors are correlations between the genes, dependence across samples, and natural variation between biological samples (Auer and Doerge, 2011).

There are mainly two ways to handle overdispersion in the literature: quasi-likelihood (QL) models and hierarchical models. First we introduce QL methods. The general form of the variance function in QL methods is $\text{Var}(Y_{mn}) = \Phi_m V_m(\mu_{mn})$, where Φ_m is a dispersion parameter that can be estimated from the data, $V_m(\mu_{mn})$ is specified by the user, and $\mu_{mn} = E(Y_{mn})$. The user-defined variance function needs to satisfy $\frac{\partial l(\mu_{mn}|y_{mn})}{\partial \mu_{mn}} = \frac{y_{mn}-\mu_{mn}}{V_m(\mu_{mn})}$, where $l(\mu_{mn}|y_{mn})$ is the corresponding QL function. Some choices of the variance function are listed in Table 9.1 in McCullagh and Nelder, 1989. Tjur has proposed to compute the test statistic $F = \frac{LRT_m/q}{\hat{\phi}_m}$ and compared it with $F(q, n-p)$ distribution, where LRT_m is the QL ratio test statistic, $\hat{\phi}_m$ is estimated by the QL models with nonlinear regression models, and q is the difference between the dimensions of the full and null-constrained mean parameter spaces (Tjur, 1998; Lund et al., 2012).

One popular method based on the two stage Poisson model (Auer and Doerge, 2011), applies the QL model by first specifying $V_m(\mu_{mn}) = \mu_{mn}$, the usual Poisson model. If there is evidence that the overdispersion parameter $\Phi_m > 1$, then Φ_m is estimated by the QL approach (Wedderburn, 1974) with the formula

$$\hat{\phi}_m = \frac{\sum_n \frac{(y_{mn}-\hat{\mu}_{mn})^2}{\hat{\mu}_{mn}}}{mn-1}, \tag{4}$$

where $\hat{\mu}_{mn}$ is the maximum likelihood estimate (MLE) of the mean μ_{mn}.

Recall, the other approach for overdispersion is hierarchical modeling. There are mainly two types of hierarchical models in the RNA-Seq data analysis literature: the beta-binomial model and the negative-binomial model.

Baggerly et al., 2003 have proposed to use beta-binomial distribution to model the data when comparing two groups. Assume that $Y_{mn}|\pi_{mn} \sim Binomial(y_{.n}, \pi_{mn})$ and $\pi_{mn} \sim Beta(\alpha, \beta)$. The model has also been adopted in Vêncio et al., 2004, and the parameters are estimated using a Bayesian method instead of frequentist methods. Zhou et al. have proposed the beta-binomial distribution with logistic regression for π_{mn} to solve the overdispersion problem for more than 2 samples (Zhou, Xia, and Wright, 2011). They assume that π_{mn} follows a Beta distribution with variance $\phi_m E(\pi_{mn})[1-E(\pi_{mn})]$, and estimate the parameters via maximum likelihood.

The multinomial distribution is also often used to model the counts in each sample. For example, in T. Wang, C. Yang, and Zhao (2019), Y is a human host phenotype based on his/her microbiota $\boldsymbol{X}$, and $\boldsymbol{z}_y = (z_{y1}, \ldots, z_{yp})^T \in \mathbb{R}^p$ is a vector of probabilities such that $\sum_{j=1}^p z_{yj} = 1$. To address the overdispersion problem, the common methods are to assume $\boldsymbol{z}_y$ is random with the Dirichlet or the additive logistic normal prior distribution. T. Wang, C. Yang, and Zhao (2019) have used the additive logistic normal multinomial distribution. They assume that

$$log(\frac{z_{yj}}{z_{yp}}) = a_j + \boldsymbol{\gamma}_j^T \boldsymbol{\beta} \boldsymbol{h}_y \tag{5}$$

for $j = 1, \ldots, p-1$, where $a_j \in \mathbb{R}, \gamma_j \in \mathbb{R}^d, \boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ is a known vector-valued function of y. By assuming that $\boldsymbol{a} = (a_1, \ldots, a_{p-1})^T \in \mathbb{R}^{p-1}$ is a realization of $\boldsymbol{A} = (A_1, \ldots, A_{p-1})^T \in \mathbb{R}^{p-1}$, and that $\boldsymbol{A}$ is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and is independent of Y, the model allows for overdispersion.

The negative binomial model can also be considered as a hierarchical model. The counts conditional on the mean $Y_{mn}|\mu_{mn}$ can be modeled as Poisson(μ_{mn}). A Poisson model can model the technical variability for some RNA-Seq data (Marioni et al., 2008). In order to account for the biological variability, the mean μ_{mn} can be modeled as a gamma distribution. The marginal probability distribution of the counts turns out to be a negative binomial distribution (Di et al., 2011). Here is another way to look at the formulation of the negative binomial model on RNA-Seq data mathematically. Assume the unknown true proportion of OTU m in sample n is π_{mn} and the conditional distribution $Y_{mn}|\pi_{mn} \sim$ Poisson($y_{.n}\pi_{mn}$). In addition, assume that π_{mn} varies between samples, but the coefficient of variation remains constant across all the samples for each OTU, which can be written as $E(\pi_{mn}) = \lambda_{mn}$ and $Var(\pi_{mn}) = \phi_m \lambda_{mn}^2$.

Based on the law of total expectation, the mean is

$$\begin{aligned} \mathrm{E}(Y_{mn}) &= \mathrm{E}[\mathrm{E}(Y_{mn}|\pi_{mn})] && (6) \\ &= \mathrm{E}[y_{.n}\pi_{mn}] && (7) \\ &= \mu_{mn}. && (8) \end{aligned}$$

Based on the law of total variance, the variance is

$$\begin{aligned} \mathrm{Var}(Y_{mn}) &= \mathrm{E}[\mathrm{Var}(Y_{mn}|\pi_{mn})] + \mathrm{Var}[\mathrm{E}(Y_{mn}|\pi_{mn})] \\ &= \mathrm{E}[y_{.n}\pi_{mn}] + \mathrm{Var}[y_{.n}\pi_{mn}] \\ &= \mu_{mn} + \phi_m \mu_{mn}^2, \end{aligned} \tag{9}$$

which is larger than the mean $\mathrm{E}(Y_{mn})$ (Chen, Lun, and Smyth, 2014; McCarthy, Chen, and Smyth, 2012). Negative binomial models along with its extensions are widely used for overdispersed RNA-Seq data.

The simplest case for negative binomial model setting is to assume the dispersion parameter to be constant across all the features. Assume $Y_{mn} \sim$

$NB(\mu_{mn}, \phi)$, then $\mathrm{E}(Y_{mn}) = \mu_{mn}$ and $\mathrm{Var}(Y_{mn}) = \mu_{mn}(1 + \mu_{mn}\phi)$. Robinson and Smyth (2008) propose a quantile adjusted conditional maximum likelihood (qCML) method to estimate the common dispersion parameter ϕ. Because it is not realistic that all the features disperse at the same level, Robinson and Smyth (2007) have proposed to use a weighted combination of common likelihood and individual likelihood to approximate an empirical Bayes solution for ϕ_m for each feature. In this way, each dispersion parameter estimate is squeezed towards the common dispersion estimate.

Instead of assuming $\sigma^2 = \mu + \phi\mu^2$, Anders and Huber (2010) have proposed to model the variance as a smooth function of the mean and estimate the variance with local regression (Anders and Huber, 2010; Di et al., 2011). They have estimated the overdispersion parameter by borrowing information across features with similar expression levels to achieve better estimates (Kvam, Liu, and Si, 2012).

The baySeq package provides an empirical Bayesian approach based on the negative binomial distribution. The details about the model setup, the estimation of prior and posterior distributions are detailed in Hardcastle and Kelly (2010).

Di et al. (2011) have proposed to use negative binomial power (NBP) distribution to model the mean-variance relationship. Assume $Y_{mn}|Z_{mn} \sim Poisson(Z_{mn})$ and $Z_{mn} \sim Gamma(\text{mean} = \mu_{mn}, \text{variance} = \phi\mu_{mn}^{\alpha})$. Then Y_{mn} follows the NBP distribution with mean μ_{mn} and variance $\mu_{mn}(1 + \phi\mu_{mn}^{\alpha-1})$. The estimates of (ϕ, α) are obtained by the conditional maximum likelihood strategy as Robinson and Smyth (2008).

McCarthy, Chen, and Smyth (2012) have proposed to use generalized linear models (GLM) to accommodate multiple covariates with the counts of each feature modeled as negative binomial distribution. They propose three categories of dispersion parameter estimation. The first one estimates a common dispersion parameter for all features by maximizing the shared likelihood function. The second one estimates a dispersion parameter as a function of the average count of each feature. The last one estimates a dispersion parameter for each feature and seeks a compromise between individual feature dispersion parameter estimator and the common dispersion parameter estimator based on Robinson and Smyth (2007).

Lund et al. (2012) proposes a method to specify the variance function in quasi-likelihood model to be negative-binomial model. Suppose $V_m(\mu_{mn}) = \mu_{mn} + \omega_m \mu_{mn}^2$ in the QL models. Then $\mathrm{Var}(Y_{mn}) = \Phi_m(\mu_{mn} + \omega_m \mu_{mn}^2)$. By using a quasi-likelihood approach based on a negative binomial distribution, the variability of dispersion parameter in the negative binomial distribution is also modeled. Three methods are proposed to estimate the dispersion parameter for the quasi-likelihood model Φ_m. The first one is deviance based approach. It is estimated as $\hat{\Phi}_m = \frac{2[l_m(\boldsymbol{y}_m|\boldsymbol{y}_m) - l_m(\hat{\boldsymbol{\mu}}_m|\boldsymbol{y}_m)]}{n-p}$, where p is the dimension of the full-model mean parameter space, and $l_m(\boldsymbol{\mu}_m|\boldsymbol{y}_m)$ is the quasi-likelihood function corresponding to the variance function chosen for feature m. The second method assumes the following relationship $d_0 \frac{\Phi_0}{\Phi_k} \sim \chi^2_{d_0}$, where d_0 is a scaling factor. Then an empirical method developed by Smyth (2004) is applied to obtain the estimates of Φ_0 and d_0. This method borrows information across features in the estimation. The third method fits a cubic spline to $log(\hat{\Phi}_m)$ versus $log(\bar{y}_{m.})$. Then the method in Smyth (2004) is applied to estimate the dispersion parameter Φ_m.

2.4.2 Statistical Challenge 2 – heterogeneous library sizes

This subsection reviews methods for heterogeneous library sizes.

The total number of reads generated in each sample are different. Recall that the differences between library sizes is attributable to biological variation and technical variation. The library sizes of repeated experiments can vary substantially even if they are run on the same sample. Comparing the counts of features across samples without considering the library sizes can cause problems. For example, if no features are differentially expressed, but sample 1 generates twice as many counts as sample 2, then direct comparison of the features between the samples may lead to the misleading conclusion that all the features are differentially expressed (J. Li et al., 2012). How can we incorporate the differences between library sizes into consideration when analyzing the data?

The simplest method is to use proportions obtained through dividing each of the counts by the corresponding library size (Robinson and Smyth, 2007; Marioni et al., 2008; Oshlack, Robinson, and Young, 2010; Kvam, Liu, and Si, 2012; T. Wang, C. Yang, and Zhao, 2019). One drawback of this method is that taking the logarithm of the proportions can lead to problems when there exists many zero counts. This can be alleviated by adding a pseudopositive arbitrary constant to the raw counts (T. Wang, C. Yang, and Zhao, 2019). The other drawback is that it can be affected by outliers and result in inaccurate normalization (Bullard et al., 2010). Let's borrow a simple example from J. Li et al. (2012) to explain the effect of outliers. Suppose that there are 101 features for two samples, for the first 100 features $Y_{m1} = 100$ and $Y_{m2} = 80$. For the 101 feature, $Y_{m1} = 0$ and $Y_{m2} = 2000$. With this method there is no need to normalize the library sizes since they are the same and it is likely that all the features can be claimed to be differentially epxressed. But it makes more sense to claim that the 101^{th} feature is differentially expressed, the rest can be normalized after excluding the last feature's counts. After normalization, the first 100 features are not claimed to be differentially expressed. To avoid the outliers' effects on the library size, some people choose the 75^{th} percentile of nonzero count distribution of each sample instead (Bullard et al., 2010; Kvam, Liu, and Si, 2012).

The NOISeq package uses counts per million (CPM) mapped reads to normalize the counts for each feature (Tarazona et al., 2011). The voom method uses the log - counts per million (log-cpm) to account for the differences between library sizes (Rapaport et al., 2013; Law et al., 2014). The transformation is made to apply the methods developed for microarray data which is continuous to RNA-Seq data which are counts (Law et al., 2014).

Since a longer feature generates more reads, when the expression level instead of abundance of reads is of interest, reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008) or the related method fragments per kilobase per million mapped reads (FPKM) (Trapnell et al., 2010; Rapaport et al., 2013; Soneson and Delorenzi, 2013) can be used to adjust for both the library size and the length of the features. However, it has been shown to introduce a bias in the per-feature variances when correcting the feature length with the RPKM method (Oshlack and Wakefield, 2009; Bullard et al., 2010; Robinson and Oshlack, 2010; Dillies et al., 2012; Rapaport et al., 2013). The FPKM method also has the same limitation. Also it has been shown that shorter features have larger variance than longer features after this type of normalization (Oshlack, Robinson, and Young, 2010; Kvam, Liu, and Si, 2012).

DESeq uses a size factor $\hat{s}_n = median_n \frac{y_{mn}}{(\prod_{n=1}^{N} y_{mn})^{1/N}}$ to adjust the library size, where the denominator is the geometric mean across samples (Anders and Huber, 2010; Dillies et al., 2012).

J. Li et al. (2012) propose to normalize the counts with factor $\frac{\sum_{n \in S} y_{mn}}{\sum_{n \in S} y_{.n}}$, where S is a set of features that are not differentially expressed. If S is taken to be the full set of features, then this is the total-count normalization proposed by Oshlack, Robinson, and Young (2010), which is not robust against outliers. The two-step procedure that leads to this normalization method is detailed in J. Li et al. (2012).

The edgeR package provides the trimmed mean of M values (TMM) method, which performs as well as the 75^{th} percentile on some RNA-Seq dataset in the literature (Robinson and Oshlack, 2010; Kvam, Liu, and Si, 2012). Define the log-fold-changes as $M_m = \log_2 \frac{Y_{mn}/Y_{.n}}{Y_{mn'}/Y_{.n'}}$ and absolute expression level as $A_m = \frac{1}{2} \log_2(\frac{Y_{mn}}{Y_{.n}} \frac{Y_{mn'}}{Y_{.n'}})$ for $Y_{m.} \neq 0$. A trimmed mean is the average after removing the upper and lower $x\%$ of the data. The TMM procedure is doubly trimmed by log-fold-changes M_{mk}^{r} (sample k relative to sample r for feature m) and by absolute intensity A_g. While the percentage of values to be trimmed can be adjusted, the default setting of edgeR package trims 30% of the M_m values and 5% of the A_g values. The normalization factor for sample k using the reference sample r is calculated as $\log_2(TMM_k^{(r)}) = \frac{\sum_{m \in G^*} w_{mk}^r M_{mk}^r}{\sum_{m \in G^*} w_{mk}^r}$, where $M_{mk}^r = \frac{log_2(Y_{mk}/Y_{.k})}{log_2(Y_{mr}/Y_{.r})}$ and $w_{mk}^r = \frac{Y_{.k}-Y_{mk}}{Y_{.k}Y_{mk}} + \frac{Y_{.r}-Y_{mr}}{Y_{.r}Y_{mr}}, Y_{mk}, Y_{mr} > 0$. By selecting one sample as a reference sample, the other trimmed mean factors and the trimmed normalization factors of other samples can be calculated and built into the statistical model.

J. Z. Li and Tibshirani, 2013 have proposed down sampling and resampling methods to adjust the difference of the library sizes between different experiments. Assume experiment 1 is the base level whose sequencing depth is 1 and define sequencing depth to be the ratio of expected values $\frac{E(Y_{.n})}{E(Y_{.1})}, 1 \leq n \leq N$. Assume the sequencing depths of the experiments are $d_1, \cdots, d_N$, and $d_{min} = min_{n=1,\cdots,N} d_n$, each count in sample n is randomly generated with probability of success $\frac{d_{min}}{d_n}$. Thus the sampled counts for feature m in experiment n has the following distribution $Y_{mn}^{'} \sim binomial(y_{mn}, \frac{d_{min}}{d_n})$, all the samples are down sampled to the same sequencing depth. Since this can be problematic when d_{min} is too small, they have proposed another method, Poisson sampling. Assume that $\bar{d} = (\prod_{n=1}^{N} d_n)^{1/N}$, and they resample the counts from each experiment using $Y_{mn}^{'} \sim Poisson(\frac{\bar{d}}{d_n} y_{mn})$. They find that when $\frac{d_{max}}{d_{min}} < 10$, down sampling and Poisson sampling have similar performance. But Poisson sampling is significantly better than down sampling when $\frac{d_{max}}{d_{min}} \geq 10$.

2.4.3 Statistical Challenge 3 – heterogeneous feature counts

This subsection introduces how heterogeneous total counts across features affect results with a focus on a method proposed by Habiger, Watts, and Anderson (2017).

The total counts for each feature are different. The expected value of the total counts for each feature is proportional to the expression level of the feature across all samples and the length of the feature. The higher the expression level, the longer the feature, the more reads are generated for the feature (Lorenz et al., 2014; Tarazona et al., 2011). Longer features are more likely to be identified when the expression levels of two features are the same (Oshlack, Robinson, and Young, 2010; Oshlack and Wakefield, 2009; E.-W. Yang, Girke, and Jiang, 2013; Young et al., 2010). This is because most statistical methods have more power when the sampling sizes are larger, which has been observed in some RNA-Seq data (Habiger, Watts, and Anderson, 2017; Oshlack and Wakefield, 2009). Currently, two methods have been proposed to deal with this challenge. A Markov random field approach, which is based on graphical models, has been proposed by E.-W. Yang, Girke, and Jiang (2013).

The other method is based on the conditional local false discovery rate ($clFDR$) proposed by Habiger, Watts, and Anderson (2017). Assume that the counts for the m^{th} feature from the n^{th} sample Y_{mn} follows a Poisson distribution Poisson(μ_{mn}) and $log(\mu_{mn}) = \alpha_m + \beta_m x_n$. When $\beta_m = 0$, μ_{mn} does not change with the quantitative trait for sample n. The larger values of $|\beta_m|$ correspond to stronger associations between the feature and the trait. By conditioning on the total counts for the feature, the nuisance parameter α_m estimation can be avoided (McCullagh and Nelder, 1989), so that $\boldsymbol{Y_m}|y_{m.} \sim Multinomial(y_{m.}, \boldsymbol{p}(\beta_m))$, where $p_n(\beta_m) = \frac{\exp\{\beta_m x_n\}}{\sum_{n=1}^{N} \exp\{\beta_m x_n\}}$. Motivated by Efron's $lFDR$ method, the counts of each feature are then modeled by a mixture of multinomial distributions. That is, assume there are $K+1$ possible values for β_m, say $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \cdots, \gamma_K)^T$ with $\gamma_0 = 0$. In addition, assume $P(\beta_k = \gamma_k) = \pi_k$ with the constraints $\sum_{k=0}^{K} \pi_k = 1$ and $0 < \pi_k < 1$. The vector of the mixing proportions is denoted as $\boldsymbol{\pi} = (\pi_0, \pi_1, \cdots, \pi_K)^T$. Define the mixture probability mass function

$$\begin{aligned} P(\boldsymbol{Y}_m|y_{m.}; \boldsymbol{\gamma}, \boldsymbol{\pi}) &= \sum_{k=0}^{K} \pi_k P(\boldsymbol{y}_m|y_{m.}; \gamma_k) && (10) \\ &= \sum_{k=0}^{K} \pi_k \frac{y_{m.}!}{\prod_{n=1}^{N} y_{mn}!} \prod_{n=1}^{N} p_n(\gamma_k)^{y_{mn}}, && (11) \end{aligned}$$

where $p_n(\gamma_k) = \frac{\exp\{\gamma_k x_n\}}{\sum_{n=1}^{N} \exp\{\gamma_k x_n\}}$.

The procedure is akin to the $lFDR$ procedure, but use the $clFDR$ instead. Define the $clFDR$ by $clFDR(\boldsymbol{y_m}, y_{m.}; \boldsymbol{\pi}, \boldsymbol{\gamma}) = \frac{\pi_0 P(\boldsymbol{y_m}|y_{m.}; \gamma_0 = 0)}{\sum_{k=0}^{K} \pi_k P(\boldsymbol{y_m}|y_{m.}; \gamma_k)}$. The

$clFDR$ procedure is detailed below:

1. Order the $clFDR$ statistics $clFDR_{(1)} \leq clFDR_{(2)} \leq \cdots clFDR_{(M)}$.

2. Find j, where $j = \max\{m : \sum_{i=1}^{m} clFDR_{(i)} \leq m\alpha\}$.

3. Reject all $H_{(m)}$, where $m = 1, 2, \cdots, j$.

4. If $\sum_{i=1}^{m} clFDR_{(i)} > m\alpha$ for all m, no hypotheses are rejected.

Habiger, Watts, and Anderson (2017) shows that by utilizing the conditional distribution, the model is more likely to detect the features which are strongly associated with the quantitative trait instead of those with large total counts. They also showed that the $clFDR$ procedure also controls FDR.

Since $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ are not known, in order to calculate the $clFDR$ statistics $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ must be estimated. This leads to the adaptive $clFDR$ procedure. Habiger, Watts, and Anderson (2017) estimates $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ with the EM algorithm.

We formally outline the details next since similar techniques are proposed in this research. Let $\boldsymbol{z}$ be a matrix where $z_{km} = 1$ if $\boldsymbol{Y_m}$ has density $P(\boldsymbol{y_m}|y_{m.}, \gamma_k)$ and 0 otherwise and denote $\boldsymbol{y}_{..} = (y_{1.}, y_{2.}, \cdots, y_{M.})^T$. The likelihood function can be written as:

$$\begin{aligned} l(\boldsymbol{\pi}, \boldsymbol{\gamma}) = P(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M|\boldsymbol{y}_{..}; \boldsymbol{\pi}, \boldsymbol{\gamma}) &= \prod_{m=1}^{M} P(\boldsymbol{y}_m|y_{m.}; \boldsymbol{\pi}, \boldsymbol{\gamma}) & (12) \\ &= \prod_{m=1}^{M} [\sum_{k=1}^{K} \pi_k P(\boldsymbol{y}_m|y_{m.}; \boldsymbol{\gamma}_k)] & (13) \\ &= \prod_{m=1}^{M} \{\sum_{k=1}^{K} \prod_{k=1}^{K} [\pi_k P(\boldsymbol{y}_m|y_{m.}; \boldsymbol{\gamma}_k)]^{z_{mk}}\}. & (14) \end{aligned}$$

The log likelihood function is

$$\log l(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \sum_{m=1}^{M} \sum_{k=1}^{K} z_{mk} [\log \pi_k + \log P(\boldsymbol{y}_m|y_{m.}; \boldsymbol{\gamma}_k)]. \tag{15}$$

The EM algorithm is iterated by expectation (E) step and maximization (M) step. Let $\boldsymbol{\pi}_k^i$ and $\boldsymbol{\gamma}_k^i$ denote values at $\boldsymbol{\pi}_k$ and $\boldsymbol{\gamma}_k$ at i^{th} iteration. The E step for the $(i+1)^{th}$ iteration is to find the conditional expectation of the z_{km} given the data $\boldsymbol{y}_m$ and current parameter values $\boldsymbol{\pi}^i$ and $\boldsymbol{\gamma}^i$

$$z_{km}^{(i+1)} = \frac{\pi_k^{(i)} f_{(i)}(\boldsymbol{y}_m | y_{m.}; \boldsymbol{\gamma}_k^{(i)})}{\sum_h \pi_h^{(i)} f_{(i)}(\boldsymbol{y}_m | y_{m.}; \boldsymbol{\gamma}_h^{(i)})}. \tag{16}$$

The M step for the $(i+1)^{th}$ iteration plugs $z_{km}^{(i+1)}$ in from (16) and maximizes the log likelihood function (16) with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ and under the constraint that $\sum_{k=0}^{K} \pi_k = 1$ via the method of Lagrange multipliers. This gives

$$\pi_k^{(i+1)} = \frac{\sum_{m=1}^{M} z_{km}^{(i+1)}}{M}. \tag{17}$$

Maximizing (16) with respect to $\boldsymbol{\gamma}$ amounts to maximizing

$$\sum_{m=1}^{M} \sum_{k=1}^{K} z_{km}^{(i+1)} \{ \sum_{n=1}^{N} y_{mn} [\gamma_k x_n - \log(\sum_{n=1}^{N} e^{\gamma_k x_n})] \}. \tag{18}$$

Any standard maximization method such as Newton-Raphson routine can be used.

Parameters are updated until $\log l(\boldsymbol{\pi}^{(i+1)}, \boldsymbol{\gamma}^{(i+1)}) - \log l(\boldsymbol{\pi}^{(i)}, \boldsymbol{\gamma}^{(i)}) < \boldsymbol{\epsilon}$.

3 Proposed Research

In this section, we propose methods that take care of all three statistical challenges mentioned in section 2.4: overdispersion, heterogeneous library sizes and heterogeneous feature counts. Analytical assessment and simulation study designs are proposed.

3.1 Proposed Method

A finite mixture of Poisson log linear models and an *lFDR* oracle procedure are proposed. Parameter estimation methods follow.

3.1.1 The Model

Assume that $Y_{mn} \sim$ Poisson(μ_{mn}) and

$$log(\mu_{mn}) = \alpha_m + \beta_m x_n + d_n, \tag{19}$$

where α_m gives us the baseline of the mean counts for the feature, β_m quantifies the association level between the mean counts and the trait, and d_n is the normalization factor to adjust the differences between library sizes.

The mean can be rewritten as $\mu_{mn} = \exp(\alpha_m + \beta_m x_n + d_n)$. Observe that when β_m and d_n are fixed, μ_{mn} increases as α_m increases. This implies that the mean counts for the m^{th} feature are positively associated with α_m for all n, and hence $y_{m.}$ is positively associated with α_m. When α_m and β_m are fixed, μ_{mn} increases as d_n increases for all m. Likewise, this implies that the library size $y_{.n}$ is associated with d_n positively. When α_m and d_n are fixed, the magnitude of β_m determines the strength of the association between feature m and the trait and the sign determines the direction of the association. Specifically, if β_m is positive, then the association is positive; if β_m is negative, then the association is negative. If $|\beta_m|$ is large, then the association is strong. When $\beta_m = 0$, there is no association between feature m and the trait.

Note that if α_m and d_n are viewed as random variables, then $Var(\alpha_m)$ and $Var(d_n)$ represent feature and library size heterogeneity. To facilitate the hypothesis testing on β_m, assume α_m and d_n are fixed for the moment and specify a prior on β_m. Assume there are $K+1$ distinct possible values for β_m, say $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \cdots, \gamma_K)^T$ with $\gamma_0 = 0$. In addition, assume $P(\beta_k = \gamma_k) = \pi_k$ with $\sum_{k=0}^{K} \pi_k = 1$ and $0 < \pi_k < 1$. Let $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_K)$ and $\boldsymbol{d} = (d_1, \ldots, d_N)$. The mixture probability mass function is defined as

$$P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{d}) = \sum_{k=0}^{K} \pi_k P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \beta_m = \gamma_k, \boldsymbol{d}), \tag{20}$$

where $P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \beta_m = \gamma_k, \boldsymbol{d}) = \prod_{n=1}^{N} P(Y_{mn} = y_{mn} | \alpha_m, \beta_m = \gamma_k, \boldsymbol{d}) =$

$\prod_{n=1}^{N} \frac{e^{\mu_{mn}} \mu_{mn}^{y_{mn}}}{y_{mn}!}$ for μ_{mn} defined as in (19).

To determine if the m^{th} feature is associated with the quantitative trait, we can test null hypothesis $H_m : \beta_m = 0$. Observe that under model (19) we have $P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \beta_m = 0, \boldsymbol{d}) = \prod_{n=1}^{N} P(Y_{mn} = y_{mn} | \alpha_m, \beta_m = 0, \boldsymbol{d}) = \prod_{n=1}^{N} \frac{e^{\mu_{mn}} \mu_{mn}^{y_{mn}}}{y_{mn}!}$ with probability π_0, where $\mu_{mn} = \exp(\alpha_m + d_n)$.

Recall, Efron (2008) shows that an empirical null hypothesis, which is a null hypothesis that can be specified using parameter estimates, can lead to a better FDR method in some settings. Under model (20), this approach can be incorporated by utilizing an estimate of γ_0 and testing null hypothesis $H_m : \beta_m = \hat{\gamma}_0$ for each m.

3.1.2 Oracle Procedure

The $lFDR$ statistic can be viewed as the posterior probability that the null hypothesis is true given the data $\boldsymbol{Y}_m = \boldsymbol{y}_m$ and all the other parameters $\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}$. Using Bayes rule, observe

$$\begin{aligned} lFDR_m &= P(\beta_m = \gamma_0 | \boldsymbol{y}_m; \alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}) &(21)\\ &= \frac{\pi_0 P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \beta_m = \gamma_0, \boldsymbol{\pi}, \boldsymbol{d})}{\sum_{k=0}^{K} \pi_k P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \beta_m = \gamma_k, \boldsymbol{d})}. &(22)\end{aligned}$$

The $lFDR$ procedure is (Sun and Cai, 2007):

1. Order the $lFDR$ statistics $lFDR_{(1)} \leq lFDR_{(2)} \leq \cdots \leq lFDR_{(M)}$.

2. Find $j = \max\{m : \sum_{i=1}^{m} lFDR_{(i)} \leq m\alpha\}$.

3. Reject all $H_{(m)}$, where $m = 1, 2, \cdots, j$.

4. If $\sum_{i=1}^{m} lFDR_{(i)} > m\alpha$ for all m, no null hypotheses are rejected.

Since $\boldsymbol{\pi}, \alpha_m, \boldsymbol{\gamma},$ and $\boldsymbol{d}$ are not known and only an oracle can know it, this procedure is called an oracle procedure.

3.1.3 Parameter Estimation

In order to implement the $lFDR$ procedure, we need to estimate the parameters and plug them into the oracle procedure or condition on row and/or column totals to avoid their estimation. We propose three ways of estimation: unconditional approach, conditional approach, and hybrid approach. The likelihood function is a function of $\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma},$ and $\boldsymbol{d}$, which is

$$l(\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}) = \prod_{m=1}^{M} P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}), \tag{23}$$

where recall $P(\boldsymbol{Y}_m = \boldsymbol{y}_m | \alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d})$ is defined as in (20).

Unconditional Approach In unconditional approach, the MLEs of $\alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma},$ and $\boldsymbol{d}$ in above equation (23) as plug-in estimates for the oracle procedure. The EM algorithm can be implemented to calculate the MLEs (Dempster, Laird, and Rubin, 1977).

Hybrid Approach We can condition on either library size to avoid estimating α_m or total feature counts to avoid estimating $\boldsymbol{d}$. For example, as per McCullagh and Nelder (1989), Habiger, Watts, and Anderson (2017) shows that $P(\boldsymbol{Y}_m | y_{m.}; \boldsymbol{\gamma}, \boldsymbol{\pi})$ is a multinomial distribution that does not depend on α_m, and hence α_m need not be estimated. However, other parameters such as $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ must still be estimated. The MLEs of all the remaining parameters can again be obtained by the EM algorithm.

Conditional Approach To avoid estimating α_m and $\boldsymbol{d}$, we can condition on both the total feature counts and the library size so that neither of them will be in the likelihood function (McCullagh and Nelder, 1989). That is, conditioning upon $y_{m.}$ and $y_{.n}$, each count Y_{mn} follows a hypergeometric distribution. Similarly, the MLEs of $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ can be obtained by the EM algorithm.

While MLEs have tractable analytical properties, they can be computationally challenging. Thus, other simpler estimates will be considered. For example, here is a list of possibilities to estimate the normalization factor d_n in the literature reviewed in Section 2.4.2.

Method	Description
Total count normalization	$\hat{d}_n = \frac{y_{mn}}{y_{.n}}$
Upper quartile (UQUA)	$\hat{d}_n = 75^{th}$ percentile of nonzero count distribution of sample n
Counts per million reads (CPM)	$\hat{d}_n = \frac{y_{mn} \times 10^6}{y_{.n}}$
Reads per kilobase per million mapped reads (RPKM)	$\hat{d}_n = \frac{y_{mn} \times 10^3 \times 10^6}{y_{.n} \times \text{feature length in bp}}$
Fragments per kilobase per million mapped reads (FPKM)	The calculation of FPKM is the same as RPKM, the difference is that FPKM is for paired-end sequencing while RPKM is for single-end RNA-Seq experiments.
A method used in DESeq	$\hat{d}_n = median_n \frac{y_{mn}}{(\prod_{n=1}^{N} y_{mn})^{\frac{1}{N}}}$
A method proposed by J. Li et al., 2012	The normalization factor is defined as $\hat{d}_n = \frac{\sum_{m \in S} y_{mn}}{\sum_{m \in S} y_{.n}}$, where S is a set of features that are not differentially expressed.
Trimmed mean of M values (TMM) from edgeR	The normalization factor for sample k using the reference sample r is calculated as $\log_2(TMM_k^{(r)}) = \frac{\sum_{m \in G^*} w_{mk}^r M_{mk}^r}{\sum_{m \in G^*} w_{mk}^r}$, where $M_{mk}^r = \frac{log_2(Y_{mk}/Y_{.k})}{log_2(Y_{mr}/Y_{.r})}$ and $w_{mk}^r = \frac{Y_{.k} - Y_{mk}}{Y_{.k} Y_{mk}} + \frac{Y_{.r} - Y_{mr}}{Y_{.r} Y_{mr}}, Y_{mk}, Y_{mr} > 0$.

3.2 Analytical Assessment

For each oracle model, overdispersion will be characterized and FDR control will be proven. Additionally, it will be shown that H_m is not rejected if β_m is negligible.

Overdispersion Consider model (19) and suppose that α_m and d_n are random variables. Observe that since $\mu_{mn} = \exp \alpha_m + \beta_m x_n + d_n$, the mean counts for feature m in sample n is

$$E(Y_{mn}) = E\{E[Y_{mn}|\alpha_m, \beta_m, d_n]\} \tag{24}$$

$$= E[\exp(\alpha_m + \beta_m x_n + d_n)] \tag{25}$$

by the law of iterated expectation. Also,

$$Var(Y_{mn}) = E\{Var[Y_{mn}|\alpha_m, \beta_m, d_n]\} + Var\{E[Y_{mn}|\alpha_m, \beta_m, d_n]\} \tag{26}$$

$$= E[\exp(\alpha_m + \beta_m x_n + d_n)] + Var[\exp(\alpha_m + \beta_m x_n + d_n)] \tag{27}$$

by the law of iterated variance. When the null hypothesis is true, i.e., α_m, β_m, and d_n are constants, $Var[\exp(\alpha_m + \beta_m x_n + d_n)] = 0$ and $E(Y_{mn}) = Var(Y_{mn})$. When α_m or β_m or d_n are variables, $Var[\exp(\alpha_m + \beta_m x_n + d_n)] > 0$ and $E(Y_{mn}) < Var(Y_{mn})$. The overdispersion thus can be modeled.

FDR control Under model (20), the oracle procedure in Section 3.1.2 controls the *FDR*. The proof can be based on Theorem 2 in Cai and Sun, 2009 and Theorem 1 in Habiger, Watts, and Anderson, 2017.

Theorem 3.1. *For $0 < \alpha \leq 1$, the oracle procedure 3.1.2 has $FDR \leq \alpha$ under the model in (20).*

Proof. Suppose the number of rejected null hypotheses $R = k$ given the whole data matrix $\boldsymbol{Y} = \boldsymbol{y}$, $E[V|\boldsymbol{Y}, R = k] = \sum_{i=1}^{k} lFDR_i \leq k\alpha$ by the construction of oracle procedure 3.1.2. By the law of iterated expectation we have

$$
\begin{aligned}
FDR &= E(\frac{V}{R}|\boldsymbol{Y}, R > 0)P(R > 0) && (28)\\
&= E[\sum_{k=1}^{M} \frac{1}{k} E(V|\boldsymbol{Y}, R = k)]P(R > 0) && (29)\\
&\leq E[\sum_{k=1}^{M} \frac{k\alpha}{k}]P(R > 0) && (30)\\
&= \alpha P(R > 0) && (31)\\
&\leq \alpha. && (32)
\end{aligned}
$$

□

Heterogeneous total feature counts Habiger, Watts, and Anderson (2017) shows that as $y_{m.} \to \infty, p_m \to 0$ for p-values calculated based on equation (3) even if β_m is only negligibly different. However, it can be shown that under model (20) $lFDR_m = P(\beta_m = 0|\boldsymbol{y}_m; \alpha_m, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{d}) \to 1$ as $y_{m.} \to \infty$ for some $\beta_m \neq 0$.

3.3 Simulation Study

The aims for this simulation study are listed.

1. Provide empirical results to assess if the three statistical challenges listed in Section 2.4 have been addressed.

2. Compare proposed methods with other existing popular methods.

3. Check the robustness of the methods under different scenarios.

4. Identify the situations when the proposed methods work well and when they do not.

Data will be generated from Poisson distributions $Y_{mn} \sim \text{Poisson}(\mu_{mn})$ satisfying model (19) and (20). Fix α_m or generate it from a distribution. When α_m is fixed, the total feature counts show no heterogeneity. When α_m is generated from a distribution, the variance of the distribution determines the heterogeneity level of the total feature counts. The data generated with small, medium, and large variance for the distribution of α_m corresponds to low, medium, and high heterogeneous levels of total feature counts. Since d_n plays a similar role as α_m, we can generate d_n in a similar manner as α_m. Habiger, Watts, and Anderson (2017) supposes that the values β_m can take on are $\boldsymbol{\gamma} = (0, 0.1, 0.3, 0.5)$ corresponding to no, low, moderate, and strong associations. They also fix $\pi_0 = 0.5$ in the simulation study. Different number of samples will be generated, say 3, 6, 9, and 12. The results can be compared to see if the number of samples affect the methods' performance. Generate the values of the trait from a uniform distribution with a reasonable range. Generate same number of features as our real data or 10,000 features as is often done in the literature, or generate different number of features for comparison.

We are going to compare the proposed methods with existing methods in the literature. Kvam, Liu, and Si, 2012 have compared three methods based on negative binomial distributions: edgeR, DESeq, and NBPSeq. Both edgeR and NBPSeq are more liberal than DESeq. Under large sample sizes, NBPSeq is more liberal than edgeR. edgeR exhibits greater sensitivity than NBPSeq in most settings, while DESeq exhibits lowest. Both edgeR and NBPSeq are poor at controlling FDR while DESeq show strong FDR control. NBPSeq often ranks truly non-differentially expressed featuers as the most differentially expressed. These three tests have been compared with TSPM in Kvam, Liu, and Si (2012) and Soneson and Delorenzi (2013). TSPM has a poor performance at controlling FDR and ranking differentially expressed genes when the sample size is small or all features are overdispersed. It performs better when the sample size gets larger and the proportion of non-overdispersed features increases.

Soneson and Delorenzi (2013) have been compared the above methods along with the following ones: baySeq, EBSeq, NOISeq, SAMSeq, ShrinkSeq, and two tests based on the empirical Bayes linear model called limma. baySeq is very conservative with good FDR control, except when sample sizes are low. EBSeq provides a liberal test with good sensitivity and poor FDR control. NOISeq is particularly adept at ranking genes when populations are differentially overdispersed, FDR control is unevaluated. ShrinkSeq exhibits high sensitivity and poor FDR control at default settings, but offers stronger FDR control with a user-controlled fold-change thresholding procedure. Both SAMSeq and the methods based on limma are conservative especially when the sample sizes are low. No genes are declared differentially expressed when only two samples per population are available. The power of SAMSeq increases as the sample size increases without high FDR. Since limma methods are designed for microarray data which is continuous and they do not address overdispersion challenge, we may not consider it.

QuasiSeq, PoissonSeq, BBSeq, BMDE, ShrinkBayes and $clFDR$ also provide R packages but have not been compared in the literature. QuasiSeq is a quasi-likelihood approach, PoissonSeq is based on Poisson distributions, BBSeq is based on beta-binomial distribution, BMDE and ShrinkBayes are full Bayesian approaches, and $clFDR$ is an empirical Bayes procedure based on multinomial distribution. BMDE and ShrinkBayes are more computationally expensive than other methods.

The following assessment metrics will be considered: the average FDR, the average power, and the average discovered effect size. The average FDR can assess how well the proposed methods control the FDR. The average power can assess how well the proposed methods can identify the features of interest. The average discovered effect size can assess if the proposed methods discover the features that are strongly associated with the trait instead of those with large total feature counts.

Glossary

biological variation The variation between different experimental units that are treated alike or observational units that are observed under the same environment. 13

counts The number of reads assigned to each feature in the library studied. 6

false discovery rate The expectation of the proportion of false discoveries among all the rejected null hypotheses. 8

false negative Also called Type II error, we fail to reject the null hypothesis when it is false. 7

false positive Also called Type I error, we reject the null hypothesis when it is true. 7

family-wise error rate The probability of the number of false positives larger than one or the probability of committing at least one Type I error. 8

library The resulting sample of DNA or cDNA fragments converted from an RNA sample, which is extracted from the biological sample. 5

library size The total number of reads for each library. 6

local false discovery rate The posterior probability of the null hypothesis being true given the data and model or test statistics. 9

overdispersion The variance of the variable is larger than the mean. 4

reads The sequence of bases from DNA strands. 6

technical replicates The library is loaded into a hollow glass slide called flow cell. We can obtain technical replicates when sequencing the same library more than once in the same flow cell. 5

technical variation The variation caused by measurement technology, which can be measured by sequencing the same sample multiple times. 13

Acronyms

FDR false discovery rate. 4, 8, 11, 22, 25, 27, 28, 30, 31

$FWER$ family-wise error rate. 8

$clFDR$ conditional local false discovery rate. 21, 22, 31

$lFDR$ local false discovery rate. 4, 9, 21, 25, 26

MLE maximum likelihood estimate. 14, 26

NGS next generation sequencing. 3

OTUs operational taxonomic units. 6

RNA ribonucleic acid. 3

RNA-Seq RNA-sequencing. 3, 13, 15

References

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10, R106. DOI: `10.1186/gb-2010-11-10-r106`. URL: `https://doi.org/10.1186/gb-2010-11-10-r106`.

Anderson, Michael and Joshua Habiger (2012). "Characterization and Identification of Productivity-Associated Rhizobacteria in Wheat". In: *Applied and Environmental Microbiology* 78.12, pp. 4434–4446.

Auer, Paul L. and Rebecca W Doerge (2011). "A Two-Stage Poisson Model for Testing RNA-Seq Data". In: *Statistical Applications in Genetics and Molecular Biology* 10.1. DOI: `10.1515/1544-6115.1826`.

Baggerly, Keith A. et al. (2003). "Differential expression in SAGE: accounting for normal between-library variation". In: *Bioinformatics* 19.12, pp. 1477–83.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Bullard, James H. et al. (2010). “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments”. In: *BMC Bioinformatics* 11, p. 94. DOI: 10.1186/1471-2105-11-94. URL: https://doi.org/10.1186/1471-2105-11-94.

Cai, T. Tony and Wenguang Sun (2009). “Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks”. In: *Journal of the American Statistical Association* 104.488, pp. 1467–1481. DOI: 10.1198/jasa.2009.tm08415. eprint: https://doi.org/10.1198/jasa.2009.tm08415. URL: https://doi.org/10.1198/jasa.2009.tm08415.

Chen, Yunshun, Aaron T.L. Lun, and Gordon K. Smyth (2014). “Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR”. In: *Statistical Analysis of Next Generation Sequencing Data*.

Chu, Yongjun and David R. Corey (2012). “RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation”. In: *Nucleic Acid Therapeutics* 22.4, pp. 271–274.

Clancy, Suzanne (2018). “RNA Functions”. In: *Nature Education* 1.1, p. 102.

Dempster, Arthur, Natalie Laird, and D.B. Rubin (1977). “Maximum Likelihood from Incomplete Data Via EM Algorithm”. In: *J. Royal Statistical Soc., Series B* 39, pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Di, Yanming et al. (2011). “The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq”. In: *Statistical Applications in Genetics and Molecular Biology* 10.1, pp. 1–28. URL: https://EconPapers.repec.org/RePEc:bpj:sagmbi:v:10:y:2011:i:1:n:24.

Dillies, Marie-Agnès et al. (2012). “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In: *Briefings in Bioinformatics* 14.6, pp. 671–683. ISSN: 1467-5463. DOI: 10.1093/bib/bbs046. eprint: http://oup.prod.sis.lan/bib/article-pdf/14/6/671/467651/bbs046.pdf. URL: https://dx.doi.org/10.1093/bib/bbs046.

Du, Ruofei and Zhide Fang (2014). “Analysis of Metagenomic Data”. In: *Statistical Analysis of Next Generation Sequencing Data.*

Efron, Bradley (2008). “Microarrays, Empirical Bayes and the Two-Groups Model”. In: *Statist. Sci.* 23.1, pp. 1–22. DOI: 10.1214/07-STS236. URL: https://doi.org/10.1214/07-STS236.

Efron, Bradley and Trevor Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* Institute of Mathematical Statistics Monographs. Cambridge University Press. DOI: 10.1017/CBO9781316576533.

Finotello, Francesca and Barbara Di Camillo (2014). “Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis”. In: *Briefings in Functional Genomics* 14.2, pp. 130–142. ISSN: 2041-2657. DOI: 10.1093/bfgp/elu035. eprint: http://oup.prod.sis.lan/bfg/article-pdf/14/2/130/715070/elu035.pdf. URL: https://doi.org/10.1093/bfgp/elu035.

Goodwin, Sara, John Mcpherson, and W Mccombie (2016). “Coming of age: Ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17, pp. 333–351. DOI: 10.1038/nrg.2016.49.

Habiger, Joshua, David Watts, and Michael Anderson (2017). “Multiple Testing with Heterogeneous Multinomial Distributions”. In: *Biometrics* 73.2, pp. 562–570.

Hardcastle, Thomas J. and Krystyna A. Kelly (2010). “baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data”. In: *BMC Bioinformatics* 11.1, p. 422. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-422. URL: https://doi.org/10.1186/1471-2105-11-422.

Kvam, Vanessa M., Peng Liu, and Yaqing Si (2012). “A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data”. In: *American Journal of Botany* 99.2, pp. 248–256.

Law, Charity W. et al. (2014). “voom: precision weights unlock linear model analysis tools for RNA-seq read counts”. In: *Genome Biology* 15.2, R29. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29. URL: https://doi.org/10.1186/gb-2014-15-2-r29.

Li, Jun Z. and Robert Tibshirani (2013). “Finding consistent patterns: a non-parametric approach for identifying differential expression in RNA-Seq data.” In: *Statistical methods in medical research* 22 5, pp. 519–36.

Li, Jun et al. (2012). “Normalization, testing, and false discovery rate estimation for RNA-sequencing data”. In: *Biostatistics* 13.3, pp. 523–538. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxr031. eprint: http://oup.prod.sis.lan/biostatistics/article-pdf/13/3/523/18610345/kxr031.pdf. URL: https://dx.doi.org/10.1093/biostatistics/kxr031.

Lorenz, Douglas J. et al. (2014). “Using RNA-seq Data to Detect Differentially Expressed Genes”. In: *Statistical Analysis of Next Generation Sequencing Data*.

Lund, Steven P. et al. (2012). “Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates”. In: *Statistical Applications in Genetics and Molecular Biology* 11.1. DOI: 10.1515/1544-6115.1826.

Marioni, John C. et al. (2008). “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays”. In: *Genome research* 18.9. DOI: 10.1101/gr.079558.108.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: *Nucleic Acids Research* 40.10. DOI: 10.1093/nar/gks042.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall. ISBN: 9780412317606. URL: http://books.google.com/books?id=h9kFH2%5C_FfBkC.

Mitra, Riten et al. (2014). “Statistical Analyses of Next Generation Sequencing Data: An Overview”. In: *Statistical Analysis of Next Generation Sequencing Data*.

Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5, pp. 621–628.

Nettleton, Dan (2014). “Design of RNA Sequencing Experiments”. In: *Statistical Analysis of Next Generation Sequencing Data*.

Oshlack, Alicia, Mark D. Robinson, and Matthew D. Young (2010). “From RNA-seq reads to differential expression results”. In: *Genome Biology* 11.12, p. 220. DOI: 10.1186/gb-2010-11-12-220. URL: https://doi.org/10.1186/gb-2010-11-12-220.

Oshlack, Alicia and Matthew J. Wakefield (2009). “Transcript length bias in RNA-seq data confounds systems biology”. In: *Biology Direct* 4.1, p. 14. ISSN: 1745-6150. DOI: 10.1186/1745-6150-4-14. URL: https://doi.org/10.1186/1745-6150-4-14.

Ozsolak, Fatih and Patrice M. Milos (2010). “RNA sequencing: advances, challenges and opportunities”. In: *Nature Reviews Genetics* 12, pp. 87–98.

Rapaport, Franck et al. (2013). “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data”. In: *Genome Biology* 14.9, p. 3158. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-9-r95. URL: https://doi.org/10.1186/gb-2013-14-9-r95.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1, pp. 139–140. DOI: 10.1093/bioinformatics/btp616. eprint: /oup/backfile/content_public/journal/bioinformatics/26/1/10.1093/bioinformatics/btp616/2/btp616.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D. and Alicia Oshlack (2010). “A scaling normalization method for differential expression analysis of RNA-seq data”. In: *Genome Biology*

11.3, R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25. URL: https://doi.org/10.1186/gb-2010-11-3-r25.

Robinson, Mark D. and Gordon K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance". In: *Bioinformatics* 23.21, pp. 2881–2887. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm453. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/23/21/2881/16860418/btm453.pdf. URL: https://dx.doi.org/10.1093/bioinformatics/btm453.

— (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data". In: *Biostatistics* 9, pp. 321–32. DOI: 10.1093/biostatistics/kxm030.

Smyth, Gordon K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments". In: *Stat Appl Genet Mol Biol* 3. DOI: 10.2202/1544-6115.1027..

Soneson, Charlotte and Mauro Delorenzi (2013). "A comparison of methods for differential expression analysis of RNA-seq data". In: *BMC Bioinformatics* 14, p. 91. DOI: 10.1186/1471-2105-14-91. URL: https://doi.org/10.1186/1471-2105-14-91.

Sun, Wenguang and T. Tony Cai (2007). "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control". In: *Journal of the American Statistical Association* 102.479, pp. 901–902.

Tarazona, Sonia et al. (2011). "Differential expression in RNA-seq: a matter of depth". In: *Genome research* 21.12, pp. 2213–2223. DOI: 10.1101/gr.124321.111. URL: https://www.ncbi.nlm.nih.gov/pubmed/21903743.

Tjur, Tue (1998). "Nonlinear Regression, Quasi Likelihood, and Overdispersion in Generalized Linear Models". In: *The American Statistician* 52.3, pp. 222–227. ISSN: 00031305. URL: http://www.jstor.org/stable/2685928.

Trapnell, Cole et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28, URL: https://doi.org/10.1038/nbt.1621.

Tringe, Susannah Green et al. (2005). "Comparative Metagenomics of Microbial Communities". In: *Science* 308.5721, pp. 554–557. ISSN: 0036-8075. DOI: 10.1126/science.1107851. eprint: http://science.sciencemag.org/content/308/5721/554.full.pdf. URL: http://science.sciencemag.org/content/308/5721/554.

Vêncio, Ricardo ZN et al. (2004). "Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE)". In: *BMC Bioinformatics* 5.1, p. 119. ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-119. URL: https://doi.org/10.1186/1471-2105-5-119.

Wang, Tao, Can Yang, and Hongyu Zhao (2019). "Prediction analysis for microbiome sequencing data". In: *Biometrics*. DOI: 10.1111/biom.13061. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13061. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13061.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10, pp. 57–63.

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method". In: *Biometrika* 61.3, pp. 439–447. ISSN: 0006-3444. DOI: 10.1093/biomet/61.3.439. eprint: http://oup.prod.sis.lan/biomet/article-pdf/61/3/439/690500/61-3-439.pdf. URL: https://dx.doi.org/10.1093/biomet/61.3.439.

Yang, Ei-Wen, Thomas Girke, and Tao Jiang (2013). "Differential gene expression analysis using coexpression and RNA-Seq data". In: *Bioinformatics* 29.17, pp. 2153–2161. DOI: 10.1093/bioinformatics/btt363. eprint: /oup/backfile/content_public/journal/bioinformatics/29/17/10.1093_bioinformatics_btt363/2/btt363.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btt363.

Ye, Yuzhen (2011). “Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment”. In: *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine* 2010, pp. 153–157. DOI: 10.1109/BIBM.2010.5706555.

Young, Matthew D. et al. (2010). “Gene ontology analysis for RNA-seq: accounting for selection bias”. In: *Genome Biology* 11.2, R14. DOI: 10.1186/gb-2010-11-2-r14. URL: https://doi.org/10.1186/gb-2010-11-2-r14.

Zhou, Yi-Hui, Kai Xia, and Fred A Wright (2011). “A powerful and flexible approach to the analysis of RNA sequence count data”. In: *Bioinformatics* 27.19, pp. 2672–2678. DOI: 10.1093/bioinformatics/btr449. URL: https://www.ncbi.nlm.nih.gov/pubmed/21810900.