PLOS ONE

RESEARCH ARTICLE

Sequence count data are poorly fit by the negative binomial distribution

Stijn Hawinkel[1]*, **J. C. W. Rayner**[2,5], **Luc Bijnens**[3,4], **Olivier Thas**[1,4,5]

[1] Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium, [2] Centre for Computer-Assisted Research in Mathematics and its Applications, School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, Australia, [3] Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson, Ghent, Belgium, [4] I-BioStat, Hasselt University, Hasselt, Belgium, [5] National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia

* stijn.hawinkel@ugent.be

OPEN ACCESS

Citation: Hawinkel S, Rayner JCW, Bijnens L, Thas O (2020) Sequence count data are poorly fit by the negative binomial distribution. PLoS ONE 15(4): e0224909. https://doi.org/10.1371/journal.pone.0224909

Editor: Shailesh Kumar, National Institute of Plant Genome Research (NIPGR), INDIA

Received: October 22, 2019

Accepted: April 6, 2020

Published: April 30, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: Stijn Hawinkel was funded by Janssen Pharmaceutical Companies of John- son and Johnson. Luc Bijnens is currently employed by Janssen Pharmaceu- tical Companies of Johnson and Johnson. The funders supervised the work and provided suggestions, but had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Sequence count data are commonly modelled using the negative binomial (NB) distribution. Several empirical studies, however, have demonstrated that methods based on the NB-assumption do not always succeed in controlling the false discovery rate (FDR) at its nominal level. In this paper, we propose a dedicated statistical goodness of fit test for the NB distribution in regression models and demonstrate that the NB-assumption is violated in many publicly available RNA-Seq and 16S rRNA microbiome datasets. The zero-inflated NB distribution was not found to give a substantially better fit. We also show that the NB-based tests perform worse on the features for which the NB-assumption was violated than on the features for which no significant deviation was detected. This gives an explanation for the poor behaviour of NB-based tests in many published evaluation studies. We conclude that non-parametric tests should be preferred over parametric methods.

Introduction

In research areas such as RNA-sequencing (RNA-Seq) and microbiomics, sequencing technologies are applied to measure the composition of mixtures of nucleic acids [1, 2]. The resulting collection of sequences is then considered as a proxy for the transcriptomic state of a tissue or cell (in RNA-Seq) or for the species composition (for the microbiome). As both research areas employ the same technologies, their data properties and analysis techniques are similar. Apart from the biological variability between samples, the multiple manipulations, going from nucleic acid extraction, reverse transcription and PCR amplification to actual sequencing, introduce additional variability into the feature count tables. It is often assumed that the sequence counts from a single feature (either a taxon or a gene) follow the *negative binomial* (NB) distribution [3, 4]. The NB distribution can be seen as a extension of the Poisson distribution that allows for overdispersion due to the biological variability. This overdispersion is also strongly related to the frequency of zeroes in the count data. As evident

Competing interests: Stijn Hawinkel was funded by Janssen Pharmaceutical Companies of Johnson and Johnson. Luc Bijnens is currently employed by Janssen Pharmaceutical Companies of Johnson and Johnson. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

from Fig 1, the overdispersion varies between features and depends on the biological nature of the samples, being notably large for microbiome data of human origin.

NB regression models [5] correct for sample-specific covariates such as *sequencing depth* or *library size*. Methods based on the NB distribution have been used for many purposes, such as clustering [6], discriminant analysis [7] and hypothesis testing [3, 8, 9]. In this paper we focus on the latter: testing for differential expression (RNA-Seq) and testing for differential abundance (microbiome). Good statistical hypothesis tests should be able to control the probability of a type I error (false positive result) at the nominal significance level and they should have sufficient power for detecting interesting biological results. Since with sequencing experiments not a single hypothesis is to be tested, but hundreds to thousands of hypotheses are tested simultaneously, these desirable properties of statistical tests can be reformulated as follows. The testing procedure (1) should be able to control the false discovery rate (FDR) at the nominal level, and (2) it should have sufficient sensitivity (i.e. true positive rate) for detecting interesting biological results.

Popular statistical tests for sequencing experiments include edgeR [10] and DESeq2 [11], which both rely on the NB assumption. These methods have been evaluated in many studies, using synthetic data generated with the very same NB distribution [6–8, 12]. As a self-fulfilling prophecy, the methods are then found to perform well, i.e. they appear to control the FDR and have good sensitivity. On the other hand, when evaluating the methods using realistically simulated data (e.g. by resampling from real datasets), comparative studies have revealed a substandard performance [13–16]. In particular, they conclude that the NB-based methods show

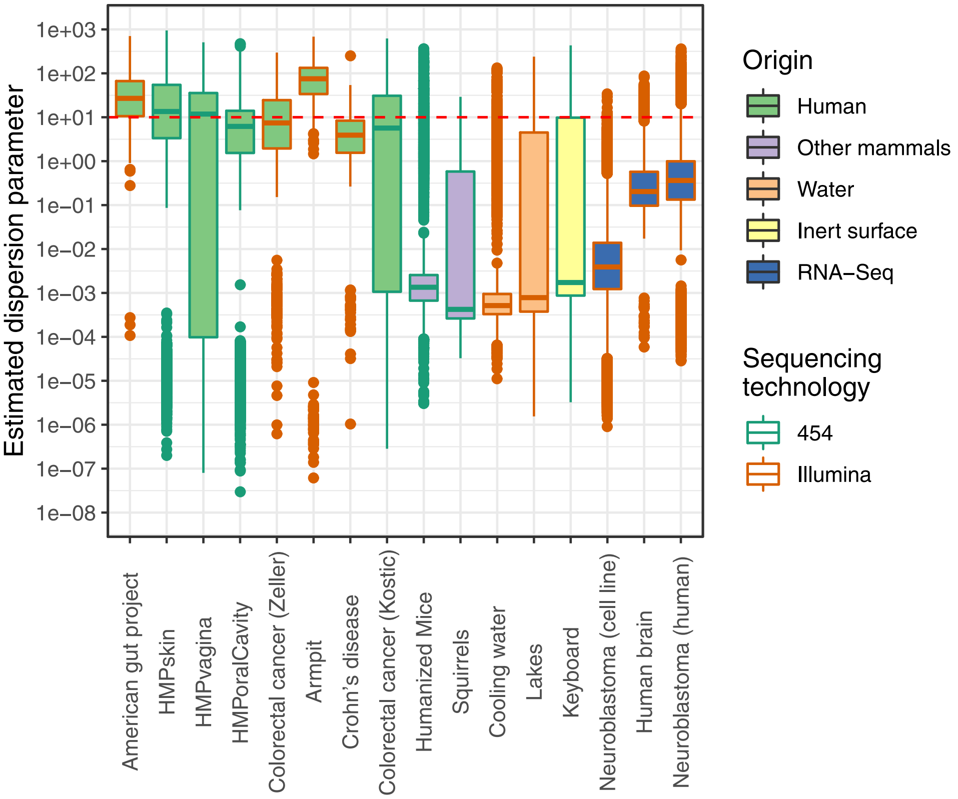

Fig 1. Boxplots of estimated feature-specific dispersion parameters of the negative binomial distribution per dataset. The red dashed line indicates the threshold above which the lack of fit to the negative binomial distribution could not be assessed reliably. See S4 Appendix for details on the datasets used.

https://doi.org/10.1371/journal.pone.0224909.g001

a poor FDR control; often the FDR is larger than the nominal level, resulting in too many false positive findings. Nonparametric methods, which do not rely on strong distributional assumptions, have been demonstrated to control the FDR with realistically simulated data, e.g. the Wilcoxon rank sum test and ALDEx2 [17].

To date no conclusive explanation has been found for this drop in performance. One obvious possible explanation is that real sequence count data are not well described by the NB distribution. The goodness of fit of ecological count data to the NB distribution has been formally tested before, but not with theoretically supported tests. These investigations only considered a limited number of datasets and they did not find a lack of fit to the NB distribution [18–20].

In this paper we propose a new statistical goodness of fit (GoF) test for the NB distribution in regression models that are commonly used for analysing RNA-Seq and microbiome studies. The performance of the new GoF test is evaluated in a simulation study. When applied to publicly available RNA-Seq and microbiome datasets, a lack of fit was discovered for many features in several datasets. Finally, the consequences of these violations to the NB assumption are investigated.

Smooth tests for the negative binomial distribution in regression models

Negative binomial regression models

NB regression models are described here for a single feature. Let (Y_i, $\boldsymbol{x}_i$), $i = 1, \ldots, n$, denote the n sample observations of outcome Y and the p-dimensional regressor $\boldsymbol{x}$. The regressor vector may include dummy variables for the coding of factor variables, and it may include a constant if an intercept is to be included in the model. We consider NB regression models of the form

$$Y \mid \boldsymbol{x} \sim \text{NB}(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}), \phi)$$

where ϕ is the overdispersion parameter and

$$\mu(\boldsymbol{x}; \boldsymbol{\beta}) = \exp(\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x}) \tag{1}$$

in which β_0 is a given constant, which is generally referred to as the offset of the regression model.

The smooth test will be constructed for the model used for comparing two experimental conditions (e.g. two treatments). This model includes an offset β_0 = log(library size), an intercept and only one regressor, x_i, defined as a 0/1 dummy variable referring to the two experimental conditions. Hence, $\boldsymbol{x}_i^t = (1, x_i)$. Model 1 thus reduces to

$$\mu(x; \beta_1, \beta_2) = \exp(\beta_0 + \beta_1 + \beta_2 x). \tag{2}$$

We further assume that, given the regressor x_i, the outcomes Y_i are independently distributed.

Construction of the test statistic

For the one-sample problem (i.e. testing the null hypothesis that a sample comes from a hypothesised distribution) many types of GoF tests have been proposed; see [21] for a comprehensive overview. Our test fits within the framework of smooth tests. For the one-sample problem this class of tests was first introduced by [22] and later generalised [23]. The main idea of smooth tests is to first embed the hypothesised distribution into a larger distribution that contains extra parameters, which we refer to as embedding parameters, such that when these extra parameters equal zero, the embedding distribution collapses to the hypothesised distribution. The GoF null hypothesis is thus equivalent to setting all embedding parameters to zero. The smooth test is then the efficient score test for this testing problem.

The extension of smooth tests from the one-sample problem to a regression setting, was first described in a PhD thesis [24]. Detailed theory and a few examples (normal, Poisson and zero-inflated Poisson regression) were recently provided [25]. In this section we give a brief outline of the construction. More details can be found in S1 Appendix.

Let $f(y;\mu, \phi)$ denote the probability mass function of the negative binomial distribution with mean μ given by Eq 2. Thus μ depends on x and on β. In particular,

$$f(y;\mu,\phi) = \mathrm{P}\{Y = y\} = \frac{\Gamma(y+\phi^{-1})}{y!\Gamma(\phi^{-1})}\left(\frac{\mu\phi}{1+\mu\phi}\right)^{y}\left(\frac{1}{1+\mu\phi}\right)^{1/\phi}, \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. Next, this distribution is embedded in a family of smooth alternatives,

$$f_J(y;\mu,\phi,\boldsymbol{\theta}) = C(\mu,\phi,\boldsymbol{\theta})\exp\left(\sum_{k=1}^{J}\theta_j h_j(y;\mu,\phi))f(y;\mu,\phi\right), \tag{4}$$

where J is the order of the smooth alternative, $\boldsymbol{\theta}^t = (\theta^1, \ldots, \theta_J)$ is the vector of embedding parameters, $C(\mu, \phi, \boldsymbol{\theta})$ is the normalisation constant, and $\{h_k(y;\mu, \phi)\}$ is a set of functions that are orthonormal on the hypothesised NB model. In particular, these functions must satisfy

$$\sum_{y=0}^{\infty} h_j(y;\mu,\phi)h_l(y;\mu,,\phi)f(y;\mu,\phi) = \delta_{jl}, \tag{5}$$

with $\delta_{jl} = 0$ if $j \neq l$ and $\delta_{jl} = 1$ if $j = l$.

The smooth test is basically the efficient score test for testing the null hypothesis H_0: $\boldsymbol{\theta} = \mathbf{0}$ within the smooth family. Since the smooth family does not only contain the embedding θ parameters, but also the parameters β and ϕ, these nuisance parameters need to be estimated from the data and the score test needs to account for their estimation. We consider the method of maximum likelihood (ML) for parameter estimation. Note that this estimation method differs from the methods used for edgeR [10] and DESeq2 [11]. However, our choice does not undermine the credibility of the results of our testing procedure: as long as the smooth test controls the type I error and has power, it is a valid testing procedure. Changing the estimation procedure, as part of the data analysis pipeline, does not alter the validity of a distributional assumption.

The next few paragraphs describe the ML estimation procedure, and shows the construction of the efficient score test statistic. For a score test we only need the ML estimators (MLE) of β and ϕ under the null hypothesis, i.e. in the hypothesised NB regression model given by probability mass function of Eq 3. First the log-likelihood function is constructed,

$$l(\beta,\phi) = \sum_{i=1}^{n} \log f(y_i;\beta,\phi).$$

Next, the score statistics for the parameters β and ϕ are computed,

$$\begin{aligned}
\frac{\partial}{\partial\beta}l(\beta,\phi) &= \sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log f(y_i;\beta,\phi) = \sum_{i=1}^{n}\frac{x_i(y_i-\mu_i)}{1+\mu_i\phi}\\
\frac{\partial}{\partial\phi}l(\beta,\phi) &= \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log f(y_i;\beta,\phi)\\
&= \sum_{i=1}^{n}\left(\Psi(y_i+\phi^{-1})-\Psi(\phi^{-1})-\log\ (1+\phi\mu_i)+1-\frac{y_i\phi+1}{1+\phi\mu_i}\right),
\end{aligned}$$

with $\Psi(\cdot)$ the digamma function. We will use the notation

$$\begin{aligned}
s_\beta(y_i;\beta,\phi) &= \frac{\partial}{\partial\beta}\log\ \ f(y_i;\beta,\phi)\\
s_\phi(y_i;\beta,\phi) &= \frac{\partial}{\partial\phi}\log f(y_i;\beta,\phi)
\end{aligned}\qquad(6)$$

for the score functions of β and ϕ, respectively.

The MLEs of β and ϕ, subsequently denoted by $\hat{\beta}$ and $\hat{\phi}$, are the solution to the system of equations $\frac{\partial}{\partial\beta}l(\beta,\phi)=0$ and $\frac{\partial}{\partial\phi}l(\beta,\phi)=0$. No analytical solution is available, and so an iterative numerical algorithm is needed.

In the presence of nuisance parameters, there generally are two approaches for the construction of efficient score tests. Either the information matrix is corrected to account for the estimation, or the orthonormal functions in the smooth family are altered to make them also orthogonal to the score functions of the nuisance parameters. We take the latter route. In particular, we require in addition to the orthonormality condition of Eq 5 that the h-functions also satisfy

$$\sum_{y=0}^{\infty}h_j(y;\mu,\phi)\boldsymbol{s}_\beta(y;\mu,\phi)f(y;\mu,\phi) = \sum_{y=0}^{\infty}h_j(y;\mu,\phi)\boldsymbol{s}_\phi(y;\mu,\phi)f(y;\mu,\phi) = 0 \qquad (7)$$

for all parameter values. For the implementation of the test in this paper, we have opted for polynomial h-functions because the resulting score test statistics have an interpretation related to deviations from the hypothesised NB distribution in terms of moments [21, 23, 26]. For example, the score test statistic based on the third order polynomial, say h_3, can detect deviations from the NB distribution in terms of the third order moment (skewness). Similarly, the statistic based on the fourth order polynomial, say h_4, detects deviations in the fourth order moment (kurtosis). In S1 Appendix it is shown how the Gram-Schmidt orthogonalisation procedure can be used for constructing the orthonormal polynomials. Note that the orthonormal constraints, expressed by Eqs 5 and 7, should hold for all μ and all ϕ. Since the former parameter equals $\exp(\beta_0 + \beta_1 + \beta_2 x)$, the orthonormal polynomials should be computed separately for the two experimental condition groups ($x = 0$ and $x = 1$), and in practice the unknown parameters should be replaced by their MLEs.

For notational comfort we denote the vector of nuisance parameters (here: $\boldsymbol{\beta}$ and ϕ) by $\boldsymbol{\eta}^t = (\boldsymbol{\beta},\phi)$. The smooth test statistic is a score test statistic which requires the score functions for the θ parameters in the smooth family of alternatives (Eq 4), evaluated under the null hypothesis. For parameter θ_k, this restricted score function is given by

$$s_{\theta,k}(y;x,\boldsymbol{\eta},\boldsymbol{\theta})|_{\boldsymbol{\theta}=\mathbf{0}} = \frac{\partial}{\partial\theta_k}\log f_J(y;x,\boldsymbol{\eta},\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\mathbf{0}} = h_k(y;x,\boldsymbol{\eta}).$$

Define

$$V_k(\boldsymbol{\eta}) = n^{-1/2}\sum_{i=1}^{n} h_k(y_i; x_i, \boldsymbol{\beta})$$

and $\boldsymbol{V}^t = (V_1(\boldsymbol{\eta}), \ldots, V_J(\boldsymbol{\eta}))$.

The score test statistic is given by

$$T_J = \boldsymbol{V}(\hat{\boldsymbol{\eta}})^t(\boldsymbol{I}^{-1}(\hat{\boldsymbol{\eta}}))_{\theta\theta}\boldsymbol{V}(\hat{\boldsymbol{\eta}}), \tag{8}$$

where $\boldsymbol{I}(\boldsymbol{\eta})$ is the Fisher information matrix for the parameter vector $(\boldsymbol{\eta}, \boldsymbol{\theta})$ in the smooth alternative, evaluated under the null hypothesis. In S1 Appendix we demonstrate that this matrix equals n times the identity matrix. This simple form is a consequence of the orthonormality conditions imposed on the orthonormal h-functions. The expression $(\boldsymbol{I}^{-1}(\hat{\boldsymbol{\eta}}))_{\theta\theta}$ in Eq 8 refers to the elements in the inverse Fisher information matrix that refers to the J-dimensional $\boldsymbol{\theta}$ vector, and hence it equals n^{-1} times the identity matrix. Thanks to this diagonal structure, we can write the test statistic T_J as a decomposition of J components, i.e.

$$T_j = V_1^2(\hat{\boldsymbol{\eta}}) + \cdots + V_J^2(\hat{\boldsymbol{\eta}}).$$

The distribution theory provided in [23] and [25] is directly applicable, resulting in the asymptotic null distributions of T_J and $\mathbf{V}(\hat{\boldsymbol{\eta}})$. Under the null hypothesis, as $n \to \infty$,

$$\boldsymbol{V}(\hat{\boldsymbol{\eta}})\xrightarrow{d}\text{MVN}(\boldsymbol{0}, \boldsymbol{I}) \qquad\qquad T_J\xrightarrow{d}\chi_J^2.$$

Since this is an asymptotic result, its practical validity for realistic sample sizes should be empirically investigated in a simulation study. This is the topic of the next section. In particular, the parametric bootstrap will be assessed. The parametric bootstrap constructs the null distribution of the test statistic based on repeatedly sampling n sample observations from the fitted hypothesised model. For the one-sample problem, the parametric bootstrap is described in detail in [23], but this procedure needs to be adjusted to the regression setting. Details are given in S2 Appendix.

Simulation study

According the theory of [25], the smooth test statistic T_J asymptotically has a χ_J^2 null distribution and the components $S_k(\hat{\boldsymbol{\eta}})$ asymptotically have standard normal distributions. Still, this does not guarantee that the use of these limiting distributions in practical settings with rather small sample sizes, gives good type I error control. In this section we empirically evaluate the null distribution of the test statistic and its p-value for realistic sample sizes. We also evaluate the tests with p-values calculated based on the parametric bootstrap procedure. For most smooth tests the bootstrap procedure gives good results [23].

We produced 6 datasets, each with counts randomly generated from NB distributions with increasing overdispersion parameter ($\phi \in \{0.01, 0.1, 1, 10, 20, 30\}$). Each dataset has $p = 500$ features in 50 samples balanced over two treatment groups. The mean of the NB distribution for feature j in sample i is given by $\log\mu_{ij} = \beta_{0i} + \beta_{1j} + \beta_2 x_i$, where x_i is a 0/1 dummy coding for the treatment group, the offsets β_{0i} (log-library sizes) are sampled from a normal distribution with mean 9.21 and standard deviation 1.15. The feature baselines β_{1j} are sampled from a uniform distribution over the interval [-13.8, -6.91], and the log-fold change β_2 is fixed to 2. These settings correspond to the parameter values observed in real datasets.

The results are shown in Fig 2. All QQ-plots of p-values from the asymptotic χ^2 approximation show a substantial deviation from the uniform distribution. This indicates that for a sample size of 50 this asymptotic approximation cannot be reliably used as it will not result in a control of the type I error rate. The QQ-plots for the bootstrap p-values, on the other hand, show good resemblance to the uniform distribution, unless the overdispersion parameter is larger than 10. Moreover, the resemblance to the uniform distribution is particularly close for small p-values, which is important for the type I error rate control.

We may conclude that the parametric bootstrap is a reliable method for p-value calculation for the tests developed in this paper, unless the overdispersion is larger than 10. The good approximation to the uniform distribution is also necessary for good FDR control in the multiple testing context. In S3 Appendix also the sampling distribution of the estimator of the overdispersion parameter is investigated. For extreme small or large values of the overdispersion parameter, bias and non-normality of the estimator is observed. This may be part of the explanation of the poor behaviour of the goodness of fit test for such extreme overdispersions.

Estimating the proportion of features with lack-of-fit

In this section, the goal is to estimate the number of features poorly fit by the NB distribution, rather than identifying which features show a deviating distribution. For this we leverage the fact that many features are being tested simultaneously, using the method developed by [27]. It relies on the property that the distribution of the p-values is a mixture distribution of p-values of features for which the null hypothesis holds, and p-values of features for which the null hypothesis does not hold. The density function of the p-values can then be described as

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_A(p),$$

where $g_0(p)$ is the density of p-values under the null distribution, $g_A(p)$ the density of p-values under the alternative distribution and π_0 the mixing proportion. The distribution $g_0(p)$ is known to be uniform on [0, 1], and the distribution $g_A(p)$ is skewed to the right. If we assume that there is some p_c for which $g_A(p|p > p_c) = 0$, we may employ the fdrtool R-package [28] for the estimation of p_c. [27] proposed a modified Grenander density estimator for $g(p)$. This estimator is similar to a nonparametric estimator of $g(p)$, with the additional restriction of being monotonically decreasing (since $g_A(p)$ is monotonically decreasing with p, and $g_0(p)$ is uniform, $g(p)$ must also be monotonically decreasing). The mixing proportion π_0 is then estimated as

$$\hat{\pi}_0 = \min\left(1, \frac{\sum_{k=1}^{p} I(p_k > p_c)}{p} \frac{1}{1 - \hat{G}(p)}\right)$$

with $\hat{G}(p)$ the distribution function corresponding to the estimate $\hat{g}(p)$. The quantity we are interested in is $1 - \pi_0$, i.e. the proportion of features that is not well fit by the NB. This approach has a very particular advantage: there is no need for very precise p-values for this estimation procedure. This means that a small number of bootstrap samples suffice, which greatly reduces the computational burden.

Finally, we note that even when a comparatively small percentage of features does not follow the NB distribution, this may still affect the analysis of an entire dataset. Typically, the negative binomial regression model is used to perform one test per feature. Next, one implements a multiple testing correction that works on the ensemble of p-values. Hence, a violation of the assumptions for some features can also affect the inference for features that *do* follow the negative binomial distribution.

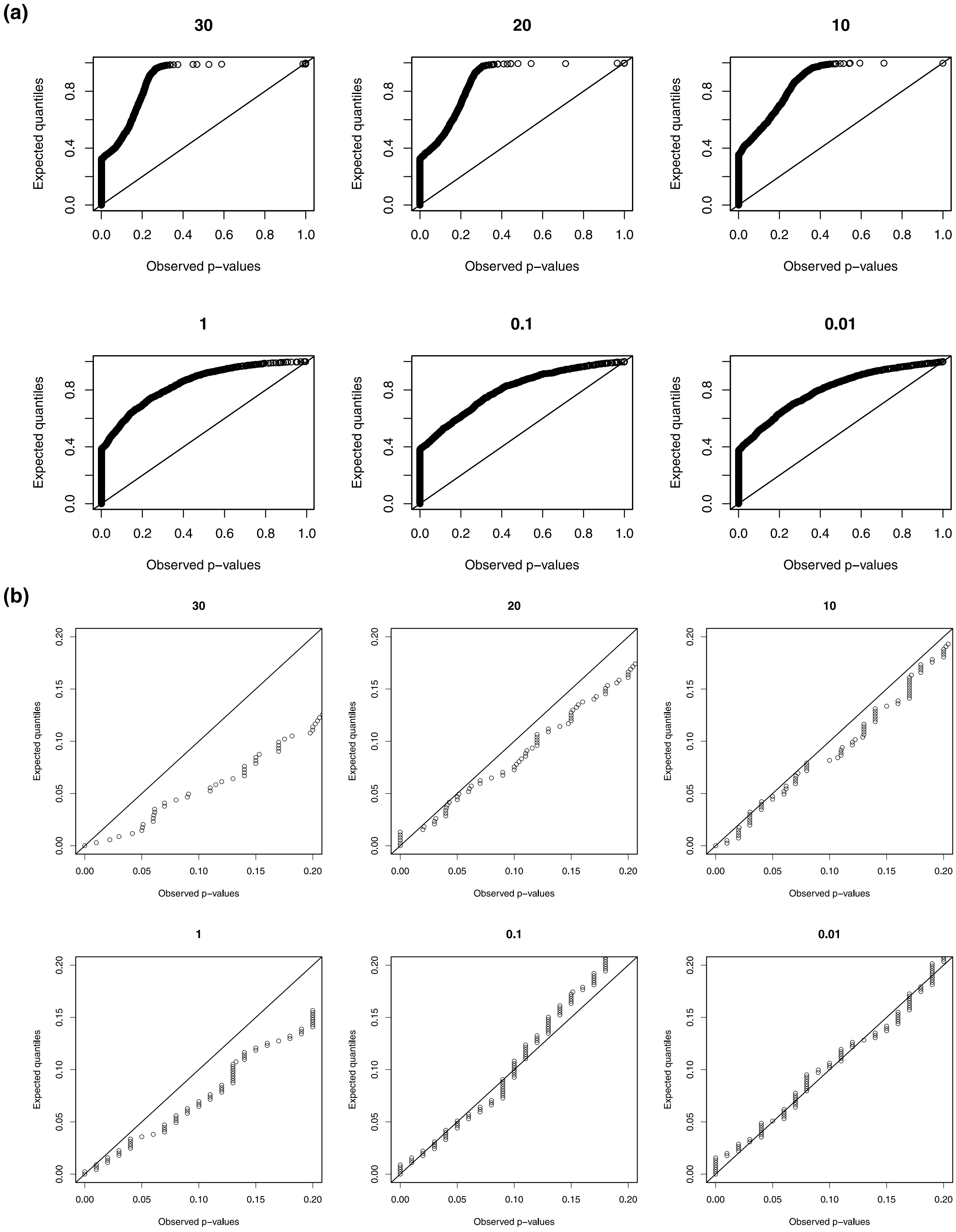

Fig 2. The uniform QQ plots of 500 p-values based on (a) the asymptotic null distribution and on (b) the bootstrap. The overdispersion parameter is printed on top of each panel.

https://doi.org/10.1371/journal.pone.0224909.g002

Application to sequencing data

The new smooth test with $J = 4$ was applied to a wide range of datasets: microbiome data from human [29–34], animal [35, 36], inert surface [37] and freshwater [38, 39] origin, as well as RNA-seq data [14, 40, 41]. More details on these datasets can be found in S4 Appendix. Features with an estimated overdispersion parameter of 10 or larger were omitted from the analysis so as to guarantee the validity of the bootstrap testing procedure (See Fig 1 for the distribution of the estimates of the overdispersion parameters). For the other features, the bootstrap (with 1,000 bootstrap runs) was used for *p*-value calculation. Subsequently, the method of the previous section was applied for estimating, for each dataset, the fraction of features poorly fit by the NB distribution. The results are summarised in Fig 3. The fraction of non-NB distributed features ranges from 2% to 90%, with many datasets having more than 25% of the features deviating from the NB assumption. For completeness we have repeated the analysis using all features; see S1 Fig. Although not all bootstrap p-values can be trusted, this analysis gives the same overall conclusion.

Zero-inflated models have been proposed as an improvement of the NB distribution [42–45]. All features were tested with a likelihood ratio (LR) test for comparing the NB against a zero-inflated negative binomial (ZINB) distribution. Based on these p-values, the fraction of significant features was estimated as before. Some exceptions notwithstanding, most of the datasets did not exhibit zero-inflation with respect to the NB distribution, or not for sufficiently many features to explain the observed lack of fit to the NB distribution (see Fig 4). Note that this LR test only serves to compare the NB and ZINB models, and does not provide

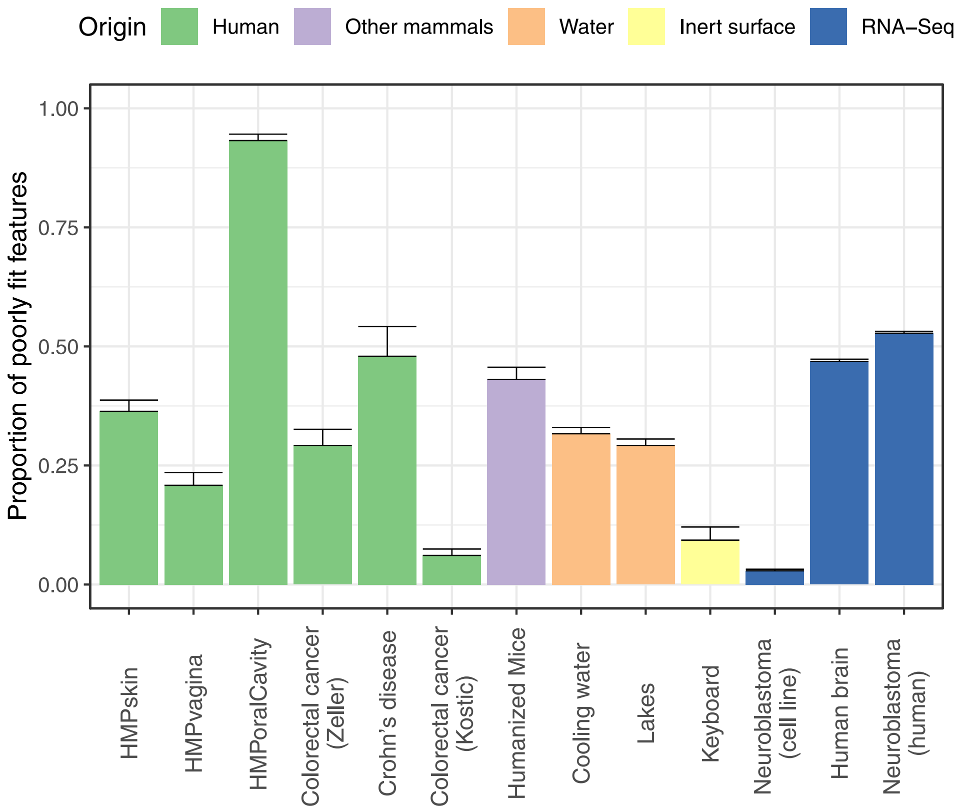

Fig 3. Estimated proportions of features with lack of fit to the negative binomial distribution per dataset. Error bars represent standard errors.

https://doi.org/10.1371/journal.pone.0224909.g003

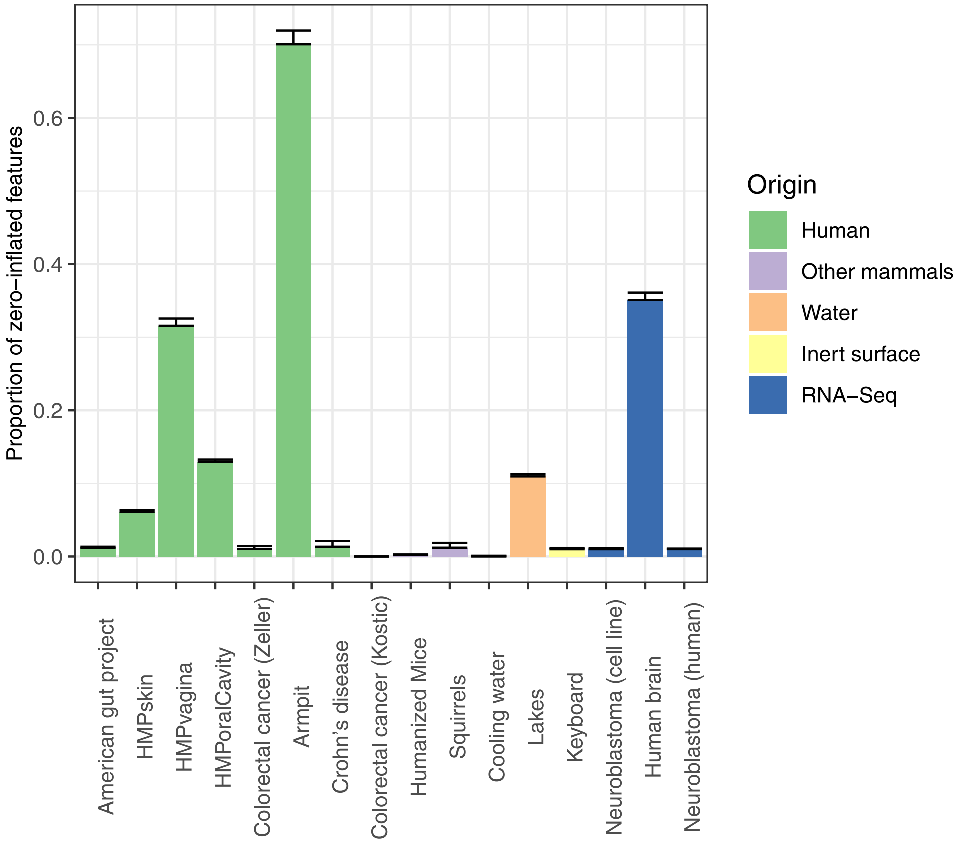

Fig 4. Estimated proportion of features with significant zero-inflation with respect to the NB distribution. Error bars represent standard errors.

https://doi.org/10.1371/journal.pone.0224909.g004

evidence that the ZINB distribution as such fits the data well. It is also important to remark that these proportions only refer to the features for which a ZINB model could be fitted and may therefore overestimate the real proportion of zero-inflated features (S1 Table shows the proportions of features without zeroes and for which hence no ZINB distribution can be fitted. Note that particularly the RNA-Seq datasets show many features without zero counts).

In S5 Appendix we further investigated possible data-related features that could explain the lack of fit to the NB assumption, but no very clear conclusions can be formulated. We observe that poorly fit features are often among the high abundant features with intermediate zero frequencies. Moreover, the fit to the NB is worse in samples with large library sizes.

Consequences

From the results from the previous section it is clear that many features in sequence count data do not follow the NB distribution. It has been shown before that the performance of statistical tests that rely on the NB distribution deteriorates when applied to realistically simulated data [13–16]. In this section we investigate the validity of NB-based hypothesis tests when applied to features that exhibited a lack of fit to the NB distribution.

Based on the bootstrap p-values of the smooth test, q-values are obtained from the procedure of [27]. All features with q-values smaller than 10% are grouped together and considered to be "poorly fit features". The other features are considered to be well fit by the NB distribution.

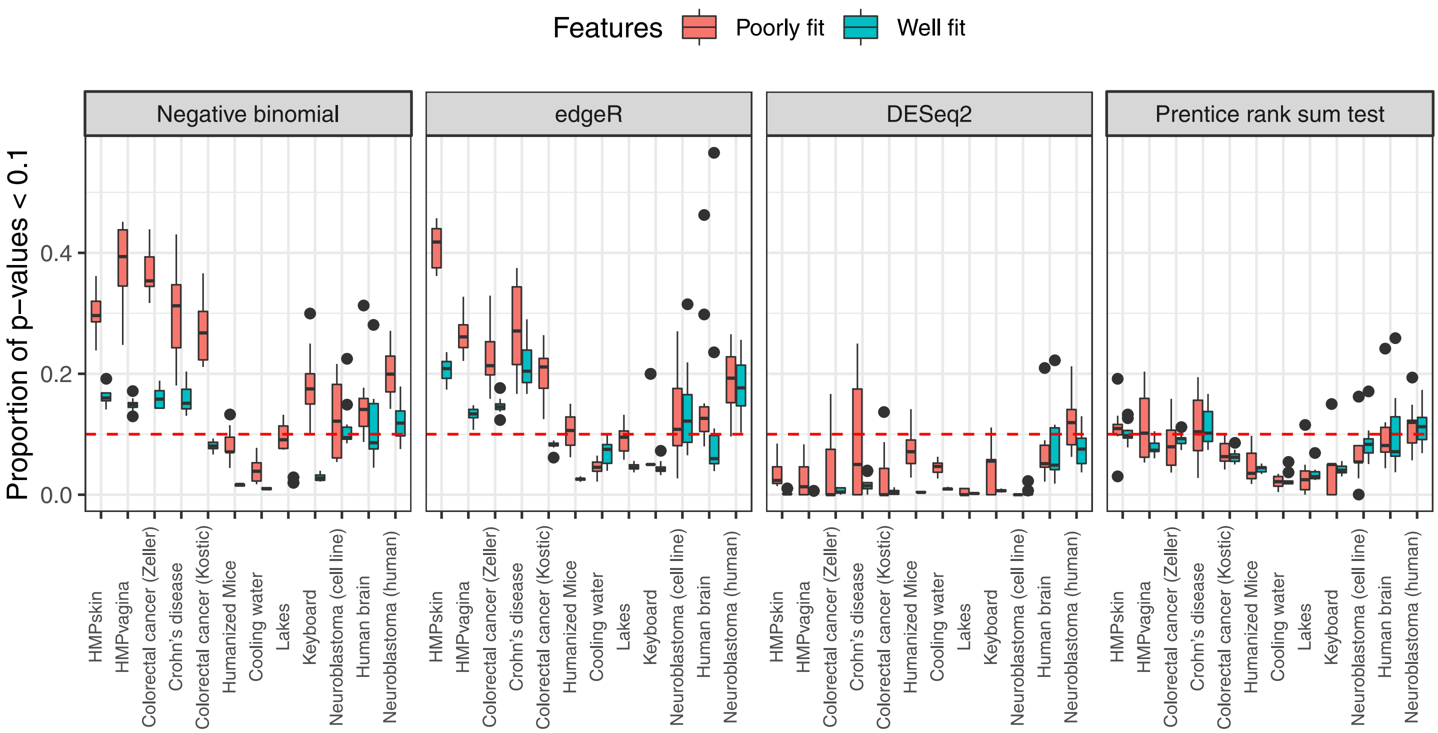

Fig 5. Boxplots of proportions of p-values smaller than 0.1 for features with poor and good fit to the NB distribution, per dataset. The p-values are computed with the Wald test in a NB model, edgeR, DESeq2 and with the Prentice rank sum test.

https://doi.org/10.1371/journal.pone.0224909.g005

For all datasets the following simulation procedure is followed. Samples are randomly allocated to one of two groups with equal probability (similar to the "real data shuffling" in [13]). Next, to every feature a negative binomial regression model is fitted with maximum likelihood, with the library sizes as offset and the original grouping variable (S4 Appendix) and the random grouping variable as regressors. P-values for the Wald test for the random grouping variable are calculated for every feature. In a similar fashion edgeR and DESeq2 were applied, as well as the nonparametric Prentice rank sum test [46]. The latter test is an extension of the Wilcoxon rank sum and Friedman tests, and it can be used for testing the effect of the random grouping variable, while accounting for the additional grouping variable. All p-values are expected to be distributed as the [0, 1] uniform distribution. This procedure is repeated 50 times for each dataset. The HMP oral cavity dataset was omitted from this study, as all features had q-values below 0.1. Fig 5 shows the proportions of p-values smaller than 0.10 for both groups of features and for all datasets. Results are shown for the four tests for differential abundance/expression. If the proportion is close to 0.1, then the test is able to control the type I error at this level. The full uniform QQ-plots (S2 Fig) are more informative, but overall the conclusions are the same.

The results show that for the Wald and edgeR tests, which both rely on the NB-distribution, the poorly fit features generally show a larger proportion of small p-values than the features for which no lack of fit to the NB was detected. This demonstrates that the poor FDR control of NB-based tests can be due to the poor fit of the NB-distribution to real sequencing count data. Overall, DESeq2 gives much better results than the other two NB-based tests. We can also see a difference depending on the origin of the datasets. For the human microbiome datasets we see large proportions, corresponding to increased type I error rates. The same conclusion holds for the RNA-Seq datasets, but to a lesser extent. The results for the environmental microbiome datasets suggest conservative type I error rate control.

For the nonparametric Prentice rank sum test, no large difference between poorly and well fit features can be observed. Moreover, the proportions are not often larger than 0.10, indicating that this test controls the type I error. This can be expected, because the test does not rely on the NB distributional assumption.

Conclusion and recommendation

Sequencing count data are often assumed to follow the NB or ZINB distributions, which form the basis of several statistical procedures for testing for differential expression (RNA-Seq) or differential abundance (microbiome). When such statistical methods are evaluated in parametric simulation studies, the count data are often generated from the same distributions. These statistical tests then seem to perform well, in the sense that they control the FDR and have good sensitivity. This is, of course, a self fulfilling prophecy. On the other hand, when these methods are evaluated in simulation studies with realistically simulated data, these NB-based methods show a poor FDR control. In particular, their true FDR is frequently much larger than the nominal level, leading to too many false positive results. The consequence of such poor FDR control, is that too many features (genes or taxa) will be called differentially expressed or abundant, potentially resulting in the publication of many false positive results, and hence contributing to the issue of poor reproducibility of scientific publications [13–16]. Nonparametric methods (the Wilcoxon rank sum test and ALDEx2), on the other hand, generally perform better in terms of controlling the FDR, but they have smaller sensitivity.

In an attempt to understand the cause of the poor FDR control of NB-based methods, we aimed to assess this distributional assumption in several public RNA-Seq and microbiome sequencing count datasets. For this purpose, we had to develop a new statistical goodness of fit test for the NB assumption in regression models, such that library size could be accounted for, and data from several experimental groups (e.g. treatments) could be simultaneously tested. This approach allows us to use more data for a single hypothesis test, hence increasing the power of the test. We developed a new smooth goodness of fit test, which was empirically evaluated in a simulation study, from which we conclude that the bootstrap version of the test gives valid results (i.e. have uniform p-value distribution under the null hypothesis) as long as the overdispersion of the NB distribution is not larger than 10.

We analysed 13 sequencing count datasets with the new goodness of fit test and we estimated the proportion of features that does not obey the NB assumption. We conclude that most datasets have more than 25% of their features deviating from the NB assumption, and 5 out of the 13 datasets have more than 40% of the features that cannot be described by the NB distribution. We also checked whether the ZINB distribution does significantly better fit to the data, but apart from 3 datasets, no considerably better fit was observed.

Finally, the features with lack of fit to the NB distributions were found to be more prone to incur false positive findings in statistical tests that rely on the NB distribution, than well fit features. On the other hand, the nonparametric rank sum test performed equally well for well fit and poorly fit features, and managed to control the type I error rate.

Researchers can thus apply our new goodness of fit test to their sequence count datasets, to see if the NB distribution provides a good fit to them. The result of this test should then guide the choice of hypothesis test, e.g. for differential abundance or expression detection. If the fit is poor, we recommend to use the nonparametric tests to analyse these data.

Supporting information

S1 Fig. Poorly fit features. Barplots of the estimated proportion of features poorly fit by the negative binomial distribution, when using all features (also those with dispersion >10). Error bars represent standard errors.
(EPS)

S2 Fig. Uniform QQ-plots. Uniform QQ-plots of the *p*-values of four tests (Wald test in a NB model, edgeR, DESeq2 and with the Prentice rank sum test) for all datasets evaluated under Mock simulations (50 simulation runs). QQ-plots for poorly and well fitted features are shown.
(EPS)

S1 Appendix. Construction of the test statistic. In this appendix more details are given about the construction of the smooth test statistic.
(PDF)

S2 Appendix. The parametric bootstrap. In this appendix more details are given about parametric bootstrap procedure for p-value calculations.
(PDF)

S3 Appendix. Estimation of the overdispersion parameter. In this appendix results of a simulation study are presented, aimed at investigating the sampling distribution of the MLE of the overdispersion parameter for a range of overdispersion values.
(PDF)

S4 Appendix. Datasets. In this appendix details are provided about the datasets used in the paper.
(PDF)

S5 Appendix. Exploration of the lack of fit. In this appendix the relationship between the fit to the NB distribution and some data-related features are graphically investigated.
(PDF)

S6 Appendix. Datasets and R-code. All datasets used in the analysis, together with the R-code used to produce the outputs.
(GZ)

S1 Table. Zero observations. Proportions of features without zero observations in all datasets under study.
(PDF)

Author Contributions

Conceptualization: J. C. W. Rayner, Olivier Thas.

Formal analysis: Stijn Hawinkel.

Methodology: J. C. W. Rayner, Olivier Thas.

Software: Stijn Hawinkel.

Supervision: Luc Bijnens, Olivier Thas.

Visualization: Stijn Hawinkel.

Writing – original draft: Stijn Hawinkel, Olivier Thas.

Writing – review & editing: J. C. W. Rayner.

References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial Analysis of Gene Expression. Science. 1995; 270(5235):484–487. https://doi.org/10.1126/science.270.5235.484 PMID: 7570003
2. Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic and Evolutionary Microbiology. 1994; 44(4):846–849. https://doi.org/10.1099/00207713-44-4-846
3. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23(21):2881–2887. https://doi.org/10.1093/bioinformatics/btm453 PMID: 17881408
4. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol. 2014; 10(4):e1003531. https://doi.org/10.1371/journal.pcbi.1003531 PMID: 24699258
5. Lawless JF. Negative binomial and mixed Poisson regression. Canadian Journal of Statistics. 1987; 15 (3):209–225. https://doi.org/10.2307/3314912
6. Di Y. Single-gene negative binomial regression models for RNA-Seq data with higher-order asymptotic inference. Stat Interface. 2015; 8(4):405–418. https://doi.org/10.4310/SII.2015.v8.n4.a1 PMID: 28042360
7. Dong K, Zhao H, Tong T, Wan X. NBLDA: Negative binomial linear discriminant analysis for RNA-Seq data. BMC Bioinformatics. 2016; 17(1):369. https://doi.org/10.1186/s12859-016-1208-1 PMID: 27623864
8. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, et al. Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics. 2017; 18(1):4. https://doi.org/10.1186/s12859-016-1441-7 PMID: 28049409
9. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11 (10):R106–R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621
10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308
11. 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. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281
12. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research. 2014. https://doi.org/10.1093/nar/gku310
13. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: Microbiome differential abundance methods do not control the false discovery rate. Briefings in Bioinformatics. 2017; p. bbx104.
14. Assefa AT, Paepe KD, Everaert C, Mestdagh P, Thas O, Vandesompele J. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. Genome Biol. 2018; 19:96. https://doi.org/10.1186/s13059-018-1466-5 PMID: 30041657
15. Benidt S, Nettleton D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. Bioinformatics. 2015; 31(13):2131–2140. https://doi.org/10.1093/bioinformatics/btv124 PMID: 25725090
16. Reeb PD, Steibel JP. Evaluating statistical analysis models for RNA sequencing experiments. Front Genet. 2013; 4:178. https://doi.org/10.3389/fgene.2013.00178 PMID: 24062766
17. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014; 2:15–15. https://doi.org/10.1186/2049-2618-2-15 PMID: 24910773
18. Mi G, Di Y, Schafer DW. Goodness-of-Fit Tests and Model Diagnostics for Negative Binomial Regression of RNA Sequencing Data. PLOS ONE. 2015; 10(3):1–16. https://doi.org/10.1371/journal.pone.0119254
19. Warton DI. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics. 2005; 16(3):275–289. https://doi.org/10.1002/env.702
20. Gierliński M, Sherstnev A, Cole C, Schurch NJ, Schofield P, Barton GJ, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015; 31 (22):3625–3630. https://doi.org/10.1093/bioinformatics/btv425 PMID: 26206307

21. Thas O. Comparing Distributions. Springer Series in Statistics. Springer New York; 2010.

22. Neyman J. 'Smooth' test for goodness of fit. Skand Aktuarietidskr. 1937; 20:150–199.

23. Rayner JCW, Thas O, Best DJ. Smooth Tests of Goodness of Fit: Using R. Wiley series in probability and statistics. Wiley; 2009.

24. Rippon P. Application of smooth tests of goodness of fit to generalized linear models; 2013. Available from: https://pdfs.semanticscholar.org/9683/bd5f6057d9f3bbf1b1f41ac8928dc7303911.pdf.

25. Rayner JCW, Rippon P, Suesse T, Thas O. Smooth Tests of Goodness of Fit for the Distributional Assumption of Regression Models. submitted;.

26. Thas O, Rayner J, Best DJ, De Boeck B. Informative statistical analyses using smooth goodness of fit tests. Journal of Statistical Theory and Practice. 2009; 3(3):705–733. https://doi.org/10.1080/15598608.2009.10411955

27. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008; 9(1):1–14. https://doi.org/10.1186/1471-2105-9-303

28. Klaus B, Strimmer K. Fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism; 2015. Available from: https://CRAN.R-project.org/package=fdrtool.

29. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. Genome Res. 2009; 19(12):2317–2323. https://doi.org/10.1101/gr.096651.109 PMID: 19819907

30. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014; 10(766). https://doi.org/10.15252/msb.20145645 PMID: 25432777

31. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012; 22(2):292–298. https://doi.org/10.1101/gr.126573.111 PMID: 22009990

32. Vandeputte D, Kathagen G, Hoe KD, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017; 551:507–EP. https://doi.org/10.1038/nature24460 PMID: 29143816

33. AmericanGut org. The American gut project. 2015;.

34. Callewaert C, Lambert J, Van de Wiele T. Towards a bacterial treatment for armpit malodour. Experimental Dermatology. 2017; 26(5):388–391. https://doi.org/10.1111/exd.13259 PMID: 27892611

35. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. Sci Transl Med. 2009; 1 (6):6ra14–6ra14. https://doi.org/10.1126/scitranslmed.3000322 PMID: 20368178

36. Carey HV, Walters WA, Knight R. Seasonal restructuring of the ground squirrel gut microbiota over the annual hibernation cycle. Am J Physiol Regul Integr Comp Physiol. 2013; 304(1):33–42. https://doi.org/10.1152/ajpregu.00387.2012

37. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci USA. 2010; 107(14):6477–6481. https://doi.org/10.1073/pnas.1000162107 PMID: 20231444

38. Props R, Schmidt ML, Heyse J, Vanderploeg HA, Boon N, Denef VJ. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. Environ Microbiol. 2018; 20(2):521–534. https://doi.org/10.1111/1462-2920.13953 PMID: 29027374

39. Props R, Kerckhof FM, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman W, et al. Absolute quantification of microbial taxon abundances. The ISME Journal. 2016; 11:584–587. https://doi.org/10.1038/ismej.2016.117 PMID: 27612291

40. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol. 2015; 16(1):133. https://doi.org/10.1186/s13059-015-0694-1 PMID: 26109056

41. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013; 45:580–EP. https://doi.org/10.1038/ng.2653

42. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013; 14(1):113–128. https://doi.org/10.1093/biostatistics/kxs031 PMID: 22988280

43. Xu L, Paterson AD, Turpin W, Xu W. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. PLoS ONE. 2015; 10(7):e0129606. https://doi.org/10.1371/journal.pone.0129606 PMID: 26148172

44. Zhang X, Mallick H, Yi N. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. Journal of Bioinformatics and Genomics. 2016;(2–2).

45. Vandenberge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert JP, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biology. 2018; 19(1):24. https://doi.org/10.1186/s13059-018-1406-4

46. Prentice MJ. On the Problem of m Incomplete Rankings. 1979; 66(1):167–170.