Journal Title Here, 2022, 1–12
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

TaxaNorm: a novel taxa-specific normalization approach for microbiome data

Ziyue Wang,[1] Dillon Lloyd,[2,3] Shanshan Zhao[1,†] and Alison Motsinger-Reif[1,†,*]

[1]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, 27709, NC, USA, [2]Departments of Biological Sciences and Statistics, North Carolina State University, 27695, NC, USA and [3]Bioinformatics Research Center, North Carolina State University, 27695, NC, USA

*Corresponding author: alison.motsinger-reif@nih.gov
†Co-senior author

Abstract

Motivation: In high-throughput sequencing studies, sequencing depth, which quantifies the total number of reads, varies across samples. Unequal sequencing depth can obscure true biological signals of interest and prevent direct comparisons between samples. To remove variability due to differential sequencing depth, taxa counts are usually normalized before downstream analysis. However, most existing normalization methods scale counts using size factors that are sample specific but not taxa specific, which can result in over- or under-correction for some taxa.
Results: We developed TaxaNorm, a novel normalization method based on a zero-inflated negative binomial model. This method assumes the effects of sequencing depth on mean and dispersion vary across taxa. Incorporating the zero-inflation part can better capture the nature of microbiome data. TaxaNorm showed improved performance compared to existing methods with both simulated and real data and can aid in data interpretation and visualization.
Availability and implementation: The 'TaxaNorm' R package is freely available for download at `https://github.com/wangziyue57/TaxaNorm` and is available from CRAN.
Contact: wangziyue57@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Key words: sequencing depth, normalization, microbiome, high-throughput sequencing

1. Introduction

There is growing evidence that microbial communities influence human health (Barcik et al., 2020; Cho and Blaser, 2012). Advanced high-throughput sequencing technologies (HTS) such as 16S ribosomal RNA (rRNA) gene amplicon sequencing and whole-genome shotgun sequencing allow researchers to survey microbial communities in a study population (Liu et al., 2021; Di Bella et al., 2013; Johnson et al., 2019; Ranjan et al., 2016; Quince et al., 2017). While HTS offers advantages in precision and accuracy, its use can be limited by sequencing depth (library size), which is the total number of reads obtained per sample from equipment (Gloor et al., 2017; Sims et al., 2014; Quinn et al., 2018). The raw data is compositional and represents only a fraction of the species abundance in each sample from an ecosystem with unknown microbial volume and thus there can be significant variation in sequencing depth between samples, even within the same biological community (Gloor et al., 2017; Lin and Peddada, 2020a; McMurdie and Holmes, 2014). Thus, the observed differential abundance (DA) between samples, which, in theory, reflects biological variation, is confounded by sequencing depth (Zaheer et al., 2018; Pereira-Marques et al., 2019; Paulson et al., 2013). Therefore, data are usually normalized to eliminate bias introduced by sequencing depth and to reflect true biological heterogeneity (McMurdie and Holmes, 2014; Weiss et al., 2017). In sufficiently normalized data, taxa abundance should be independent of sequencing depth across samples.

Normalization approaches currently used for microbiome data can be broadly classified into three types - rarefaction, log-ratio transformation, and scaling. Rarefaction is commonly used in early-stage microbiome studies (Hughes and Hellmann, 2005; Navas-Molina et al., 2013; Koren et al., 2013). Reads are randomly drawn without replacement in each sample, such that all samples have

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

the same total count and thus the same sequence depth (Hughes and Hellmann, 2005). A major limitation of rarefaction is the use of an arbitrary cut-off value across samples, resulting in a loss of statistical power and sample heterogeneity due to decreased sample size (McMurdie and Holmes, 2014). Log-ratio transformation is used to normalize compositional data by taking the log-ratios of all taxa with respect to a fixed reference component (Aitchison, 1982; Egozcue et al., 2003; Mandal et al., 2015; Morton et al., 2019; Fernandes et al., 2014). To handle the zeros commonly seen in microbiome data (Paulson et al., 2013), an arbitrary pseudo count is typically used to replace zeros, but the choice of this arbitrary value can influence downstream analysis (Lin and Peddada, 2020b; Costea et al., 2014; Paulson et al., 2014). Further, the statistical inference is based on relative change with respect to the chosen reference. In order to recover the true scale of taxa and compare differences in the absolute counts, we focus on scaling in this manuscript.

Scaling is a common normalization approach that divides raw counts by a sample-specific size factor across all taxa. Algorithms to estimate size factors include total-sum scaling (TSS), which simply scales samples by their sequencing depth, median-by-ratio (MED) from DESeq2 (Love et al., 2014), upper quartile (UQ) (Bullard et al., 2010) and trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010) from edgeR (Robinson et al., 2010), cumulative sum scaling (CSS) from metagenomeSeq (Paulson et al., 2013), Wrench (Kumar et al., 2018), and analysis of compositions of microbiomes with bias correction (ANCOM-BC) (Lin and Peddada, 2020a). A major drawback of most scaling methods is the use of a common size factor to represent sequencing efficiency, which is the effect of sequencing depth on all taxa in a given sample. In practice, however, there is evidence that sequencing efficiency varies across taxa, as a particular taxon may be preferentially measured during sequencing due to polymerase chain reaction amplification efficiency or other technical reasons (Gonzalez et al., 2012; Wu et al., 2010; v. Wintzingerode et al., 1997; Paulson et al., 2013; McLaren et al., 2019). For example, sequencing efficiency varies for gram-negative and gram-positive bacteria because it is more difficult to extract DNA from gram-positive bacteria during sample collection due to their strong cell walls. As a result, gram-positive bacteria may be under-represented in observed taxa abundance (Lin et al., 2022).

Figure 1 provides an example from a human gut microbiome dataset of 510 taxa from 300 healthy individuals described in detail in Section 4.2. We examined the relationship between the counts and sequencing depth for each taxon. Two specific taxa, *Dehalobacterium* and *Bacteroides*, were shown as an example. The monotonic increasing trend in the raw counts differed (Figure 1a), and neither TSS nor ANCOM BC had good performance as they led to over- or under-correction of the effect of sequencing depth (Figure 1b, 1c). We further regressed all taxa counts on sequencing depth individually from the above data using zero-inflated negative binomial (ZINB) regression and compared the corresponding coefficients for each taxon. Under the assumption of scaling methods, coefficients should be fixed and nearly identical. To better visualize the results, we combined the taxon-specific coefficients by their phylum group. However, density plots indicate that coefficients differ by taxa (Figure 2), suggesting that the relationship between taxa count and sequencing depth varies across taxa.

Additionally, high heterogeneity has been observed within samples in microbiome data (Chen et al., 2016; Scher et al., 2013). Although CSS and Wrench consider a taxon-specific factor, they rely on zero-inflated Gaussian distribution where the sample heterogeneity will be missed. Consequently, the choice of normalization method should be not only sample-specific but also taxon-specific and based on a model that adequately captures the nature of microbiome data.

Motivated by these observations, we developed TaxaNorm, a novel normalization method based on a ZINB model, that allows the effects of sequencing depth on taxa abundance to vary by microbial organism. TaxaNorm can handle both structural (biological) and sampling zeros. Further, this method allows the magnitude of over-dispersion to depend on sequencing depth for microbiome data. In contrast to traditional fixed-dispersion negative binomial models such as DESeq2 and edgeR, TaxaNorm includes a sequencing depth-dependent dispersion parameter to account for sample heterogeneity. The output from TaxaNorm can be used for variable selection, dimensional reduction, clustering, visualization, and differential abundance analysis.

2. Materials and methods

2.1. Varying-dispersion zero-inflated negative binomial model

We assume the observed taxa counts follow a ZINB distribution, which is a mixture of a negative binomial (NB) distribution of counts and a mass distribution at zero. The excess of zeros in microbiome data are handled in two ways: structural (biological) zeros through the mass distribution at zero and sampling zeros through the NB distribution. For a given taxon i ($i = 1, ..., p$) in sample j ($j = 1, ..., n$), let $Y_{ij} \sim \mathbf{ZINB}(\mu_{ij}, \theta_{ij}, \pi_{ij})$ denote the observed counts so that

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } \pi_{ij} \\ \mathbf{NB}(\mu_{ij}, \theta_{ij}), & \text{with probability } 1 - \pi_{ij}, \end{cases} \tag{1}$$

where μ_{ij} and θ_{ij} are the mean and dispersion of the NB distribution and π_{ij} is the probability of zero mass, or the called zero-inflation parameter. Under this parameterization, the variance of NB distribution is $\sigma^2_{ij} = \mu_{ij} + \frac{\mu^2_{ij}}{\theta_{ij}}$. In particular, this NB distribution converges to a Poisson distribution when $\theta_{ij} \to \infty$.

To account for both sample- and taxon-specific effects of sequencing depth on counts, for a given taxon i, we have

$$\begin{aligned} \log(\mu_{ij}) &= \beta_{i,0} + \beta_{i,1} X_j \\ \log(\theta_{ij}) &= \kappa_{i,0} + \kappa_{i,1} X_j \\ \log(\frac{\pi_{ij}}{1 - \pi_{ij}}) &= \gamma_i, \end{aligned} \tag{2}$$

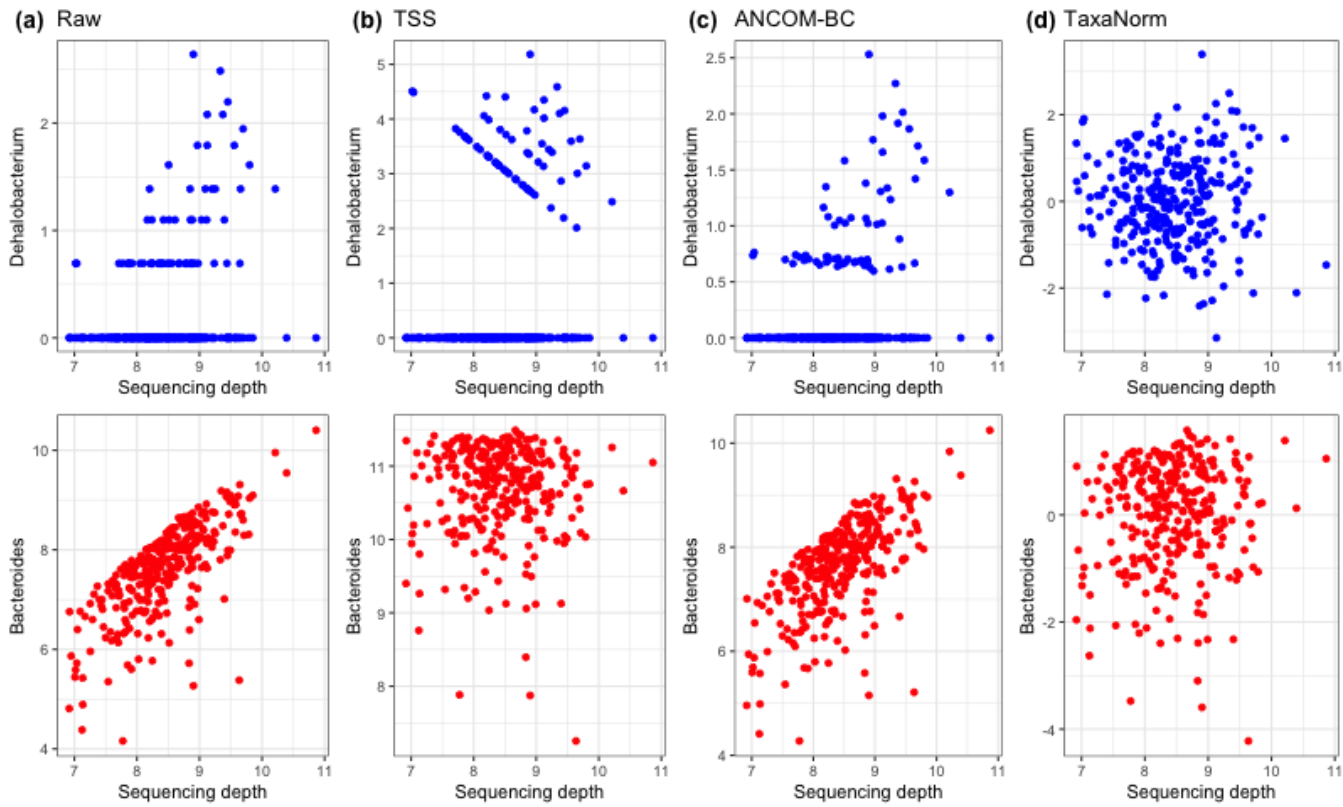

Fig. 1: Relationship between counts and sequencing depth for two microbial organisms, *Dehalobacterium* (blue) and *Bacteroides* (red). (a) Raw counts. (b) Normalized counts by TSS. (c) Normalized counts by ANCOM-BC. (d) Normalized counts by TaxaNorm. Each stool sample is represented by a single dot. The counts and depth are shown in log-scale.

where $X_j = \log(\sum_i y_{ij})$ is the log of sequencing depth for sample $j = 1, ..., n$. This formulation allows for the taxon-specific impact of sequencing depth on mean count ($\beta_{i,1}$) and dispersion ($\gamma_{i,1}$) to better capture the between-taxa variation and high heterogeneity in microbiome data compared to existing methods. Although many experimental and biological factors are linked with true zeros, we assume the zero-inflation parameter π_{ij} is taxon-specific only and is common across samples (j) since numerical evidence shows simpler models tend to have better model-fitting performance (Jiang et al., 2023; Silverman et al., 2020). Under the special case that $\beta_{i,1}$ $(i = 1, ..., p)$ is equal to 1 for all taxa and $\kappa_{i,1} = 0$ $(i = 1, ..., p)$, TaxaNorm operates under a similar model assumption as most scaling normalization methods. On the other hand, when $\kappa_{i,1} = 0$ $(i = 1, ..., p)$, this is similar to the concept behind CSS and Wrench that dispersion does not change with sequencing depth.

2.2. Parameter estimation

We fit the model and estimate the parameters for each taxon individually. For taxon i, given the observed counts $Y_i = \{y_{ij}, j = 1, ..., n\}$, and the sequencing depth $X = \{x_j, j = 1, ..., n\}$, the log-likelihood can be written as follows:

$$l(\Lambda; Y_i) = \sum_{j=1}^{n} \log \left\{ \pi_{ij} \mathbf{I}(y_{ij} = 0) + (1 - \pi_{ij}) \frac{\Gamma(y_{ij} + \theta_{ij})}{\Gamma(y_{ij} + 1)\Gamma(\theta_{ij})} \left(\frac{\mu_{ij}}{\theta_{ij} + \mu_{ij}}\right)^{y_{ij}} \left(\frac{\theta_{ij}}{\theta_{ij} + \mu_{ij}}\right)^{\theta_{ij}} \right\}, \tag{3}$$

where $\Lambda = (\beta_{i,0}, \beta_{i,1}, \gamma_i, \kappa_{i,0}, \kappa_{i,1})$ denote the full set of unknown parameters in (2), and $\mathbf{I}(\cdot)$ is an indicator function.

In practice, directly maximizing (3) causes difficulty when distinguishing zeros from the NB part and the zero-inflation part, leading to an unreasonably low estimation of π_{ij} (Jiang et al., 2023). Therefore, we used the expectation-maximization (EM) algorithm to obtain the maximum likelihood estimations (MLEs) $\hat{\Lambda} = (\hat{\beta}_{i,0}, \hat{\beta}_{i,1}, \hat{\gamma}_i, \hat{\kappa}_{i,0}, \hat{\kappa}_{i,1})$. We defined a latent random variable, Z_{ij}, to indicate whether Y_{ij} is generated from the zero mass ($Z_{ij} = 1$) or NB count ($Z_{ij} = 0$). The log-likelihood now becomes

$$l_c(\Lambda; Y_i, Z) = \sum_{j=1}^{n} \left\{ z_{ij} \log \pi_{ij}(\Lambda) + (1 - z_{ij}) \log (1 - \pi_{ij}(\Lambda)) + (1 - z_{ij}) \log f_{nb}(y_{ij}; \mu_{ij}(\Lambda), \theta_{ij}(\Lambda)) \right\}, \tag{4}$$

where f_{nb} denotes the probability mass function (PMF) of the NB distribution.

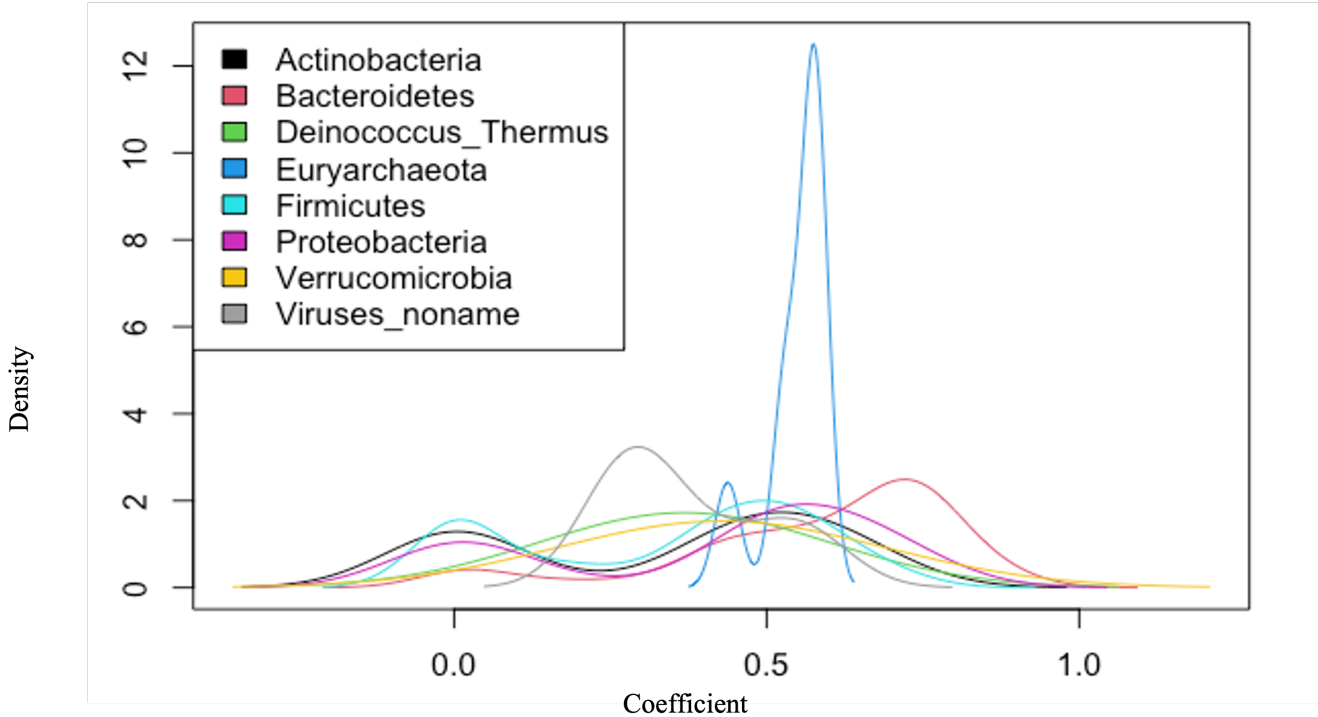

Fig. 2: Sequencing efficiency of each taxon in human gut microbiota data. The relationship between raw taxa-specific counts and sequencing depth was estimated using ZINB regression. Densities of corresponding coefficients are colored by phylum rank for all taxa.

We set the starting values $(\hat{\beta}_{i,0}^{(0)}, \hat{\beta}_{i,1}^{(0)}, \hat{\kappa}_{i,1}^{(0)})$ at the estimates from a ZINB regression with $\kappa_{i,1} = 0$, using the R built-in function *zeroinfl()* from the *pscl* package. For $\hat{\gamma}_i^{(0)}$, we initialized from a logistic regression with $Y_i = 0$ as the outcome to avoid the local maximum at the starting point (Jiang et al., 2023).

E step. For the tth iteration, the conditional expectation of the log-likelihood given the observed data Y_i and current parameter estimate $\hat{\Lambda}^{(t)}$ is computed as:

$$\begin{aligned} Q\left(\Lambda, \hat{\Lambda}^{(t)}\right) =& \mathrm{E}\left\{l_c(\Lambda; Y_i, Z)|Y_i, \hat{\Lambda}^{(t)}\right\} \\ =& \sum_{j=1}^{n} \left\{w_{ij}^{(t)} \log \pi_{ij}(\Lambda) + \left(1 - w_{ij}^{(t)}\right) \log\left(1 - \pi_{ij}(\Lambda)\right) + \left(1 - w_{ij}^{(t)}\right) \log f_{nb}\left(y_{ij}; \mu_{ij}(\Lambda), \theta_{ij}(\Lambda)\right)\right\} \end{aligned} \tag{5}$$

where

$$\begin{aligned} w_{ij}^{(t)} &= \mathrm{P}\left(z_{ij} = 1 \mid y_{ij}, \hat{\Lambda}^{(t)}\right) \\ &= \frac{\pi_{ij}\left(\hat{\Lambda}^{(t)}\right) \mathbf{I}\left(y_{ij} = 0\right)}{\pi_{ij}\left(\hat{\Lambda}^{(t)}\right) \mathbf{I}\left(y_{ij} = 0\right) + \left(1 - \pi_{ij}\left(\hat{\Lambda}^{(t)}\right)\right) f_{nb}\left(y_{ij}; \mu_{ij}\left(\hat{\Lambda}^{(t)}\right), \theta_{ij}\left(\hat{\Lambda}^{(t)}\right)\right)}. \end{aligned} \tag{6}$$

M Step. The parameter estimate is updated as $\hat{\Lambda}^{(t+1)}$, maximizing the quantity $Q\left(\Lambda, \hat{\Lambda}^{(t)}\right)$:

$$\hat{\Lambda}^{(t+1)} = \arg\max_{\Lambda} Q(\Lambda, \hat{\Lambda}^{(t)}). \tag{7}$$

These two steps are repeated until convergence is achieved:

$$\frac{\left|Q\left(\Lambda, \hat{\Lambda}^{(t+1)}\right) - Q\left(\Lambda, \hat{\Lambda}^{(t)}\right)\right|}{\left|Q\left(\Lambda, \hat{\Lambda}^{(t)}\right)\right|} < \epsilon, \tag{8}$$

where ϵ is a small value threshold (here, $1e-5$).

Finally, the estimated $\hat{\mu}_{ij}$, $\hat{\theta}_{ij}$, and $\hat{\pi}_{ij}$ are calculated by plugging in $\hat{\Lambda}$ to the above regression model (2). If the data are not truly zero-inflated, using ZINB can result in less power. Accordingly, in practice, we incorporate a pre-possessing step to divide taxa according to their zero counts. For those with less than 5% zero counts, we instead use the varying-dispersion NB regression model to estimate $\hat{\mu}_{ij}$ and $\hat{\theta}_{ij}$, by forcing $\pi_{ij} = 0$ (see details in the Supplementary Material). Kaul et al. (2017) proposed a more sophisticated method based on differentiate structure zeros that we plan to include in our package in the future.

2.3. Taxa-specific normalization and goodness-of-fit testing

Using the MLEs ($\hat{\mu}_{ij}$, $\hat{\theta}_{ij}$, $\hat{\pi}_{ij}$), we calculated the quantile residuals (Dunn and Smyth, 1996) as normalized taxa counts to remove the effects of sequencing depth:

$$y_{ij}^{\text{norm}} = \Phi^{-1}(F_{zinb}(y_{ij} - 1; \hat{\mu}_{ij}, \hat{\theta}_{ij}, \hat{\pi}_{ij}) + u_{ij} \cdot f_{zinb}(y_{ij}; \hat{\mu}_{ij}, \hat{\theta}_{ij}, \hat{\pi}_{ij})), \tag{9}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution, F_{zinb} and f_{zinb} denote the CDF and PMF of the ZINB distribution, and u_{ij} is a random variable from a uniform distribution on $(0, 1)$. Positive residuals for a given taxon in a given sample indicate greater abundance than expected given the taxon's average abundance in the microbial community and sequencing depth while negative residuals indicate less abundance.

Correctly specifying the model is critical to increase power for any analysis. Therefore, we propose two model diagnostic tests. First, we test the existence of a sequencing depth effect on taxa counts from two perspectives - mean and dispersion (the 'prevalence' test):

$$\mathbf{H}_0 : \beta_{i,1} = \kappa_{i,1} = 0, \forall i = 1, ..., p, \tag{10}$$

where the alternative hypothesis ($\mathbf{H}_A$) indicates that a sequencing depth effect exists for at least one taxon through one parameter. To do so, we use the likelihood ratio test (LRT). Specifically, under the null hypothesis, we fit a reduced intercept-only ZINB regression with fixed dispersion, where sequencing depth does not influence the taxa abundance. For this global test for all taxa, the likelihood ratio statistic is asymptotically, χ^2, distributed with a degree of freedom of $2p$.

Additionally, two major improvements of TaxaNorm are that it assumes the effect of sequencing depth is taxon-specific (i.e., differential sequencing efficiency), and the dispersion parameter depends on sequencing depth, while most scaling methods basically assume $\beta_{1,1} = ... = 1, \kappa_{1,1} = ... = 0$. To test whether TaxaNorm better reflects the data than existing scaling methods, we conduct an 'equivalence' test with the following null hypothesis

$$\begin{aligned} \mathbf{H}_0 : \beta_{1,1} = ... = \beta_{p,1} = 1, \\ \kappa_{1,1} = ... = \kappa_{p,1} = 0. \end{aligned} \tag{11}$$

Again, we used LRT where the MLE under the null hypothesis was estimated by restricting the equal effect of sequencing depth to be 1 on taxa abundance via the mean parameter, and fit a fixed dispersion ZINB regression. In this case, the likelihood ratio statistic is also asymptotically, χ^2, distributed with a degree of freedom of $2p$.

3. Simulation studies

For all simulations, we set the true values of regression parameters as those estimated from a real microbiota dataset in a subset of 147 stool samples from the human microbiome study detailed in Section 4.2.1, which ensured our simulated data have similar characteristics as actual case-study data. The dataset comprised 510 taxa with non-zero counts observed in more than 10 samples.

For a given sample size n, we first randomly generated sequencing depth, $X_j \sim \mathbf{U}(a, b)$, $j = 1, 2, ..., n$, where a and b are the minimum and maximum sequencing depths in the template data. Together with the coefficients estimated from the template data ($\hat{\beta}_{i,0}, \hat{\beta}_{i,1}, \hat{\gamma}_i, \hat{\kappa}_{i,0}, \hat{\kappa}_{i,1}$), we calculated the mean $\hat{\mu}_{ij}$, dispersion $\hat{\theta}_{ij}$, and zero mass $\hat{\pi}_{ij}$ for each taxon across all samples based on model (2). We then generated taxa counts from ZINB distribution with these mean, dispersion, and zero mass parameters.

Parameter estimation and diagnosis tests. We first assessed the performance of the proposed model diagnosis tests by considering scenarios with various coefficient settings. Specifically, we simulated under the two null hypothesis 1) $\beta_{i,1} = \kappa_{i,1} = 0$ for all taxa, where sequencing depth effect does not exist for any taxa; and 2) $\beta = 1$, $\kappa = 0$, where the sequencing depth effect is the same across all taxa with fixed dispersion, to assess the type I error control for the tests proposed in (10) and (11). We conducted power analysis for prevalence tests by simulating various alternatives, where $\beta_{i,1} = \kappa_{i,1} = 0$ for a subset of taxa, while other parameters came from real data, so the sequencing depth effect exists for several taxa only. We also conducted power analysis for equivalence tests by restricting $\beta = 1$ for a subset of taxa, while other parameter values varied. We set the sample size from 100 to 1000 for all scenarios, and repeated each simulation scenario procedure 1000 times. See Supplementary Table S1 for details of the full simulation settings.

DA taxa comparison. One way to evaluate the efficiency of data normalization is by assessing the influence of a method in downstream analysis. We conducted additional simulations under various settings to examine the performance of our proposed normalization method in detecting DA taxa using post-normalization counts.

To introduce DA taxa, we randomly assigned half of the samples to each group and manipulated the mean taxa count by multiplying the effect size. Specifically, we used a dummy variable, $G_j = \{ \begin{matrix} 0, & j \in \textbf{group 1} \\ 1, & j \in \textbf{group 2} \end{matrix}$, such that $\log(\mu_{i,G_j}) = \beta_{i,0} + \beta_{i,1}X_j + \beta_{i,2}G_j$. Here, $\beta_{i,2}$ can be regarded as the log fold-changes (i.e., $\log(\frac{\mu_{i,j\in\textbf{group 2}}}{\mu_{i,j\in\textbf{group 1}}}) = \beta_{i,2}$). For example, we first allowed 50% of taxa to be DA, by

Table 1. Power and type I error for model diagnosis test.

Sample size	Prevalence test: % of taxa without sequencing depth effect					Equivalence test: % of taxa with non-equal sequencing depth effect				
	type I error	power				type I error	power			
	100%	95%	90%	80%	50%	100%	95%	90%	80%	50%
100	0.010	0.007	0.015	0.091	0.619	0.001	0.988	0.982	0.983	0.992
200	0.013	0.31	0.829	0.984	1	0.060	1	1	1	1
500	0.010	0.999	1	1	1	0.086	1	1	1	1
1000	0.003	1	1	1	1	0.071	1	1	1	1

randomly selecting 25% of all taxa with increased abundance in group 2 with the fold-changes ($\exp(\beta_{i,2})$) generated from $\mathbf{U}(4, 10)$, and another 25% of randomly selected taxa with decreased abundance in group 2 (i.e., $\frac{1}{\exp(\beta_{i,2})} \sim \mathbf{U}(4, 10)$). The remaining 50% of taxa were not differentially abundant (i.e., $\beta_{i,2} = 0$). We also investigated scenarios with varying percentages of DA taxa, namely 10% (5% increase and 5% decrease) and 20% (10% increase and 10% decrease), as well as smaller fold changes generated from $\mathbf{U}(2, 4)$. We varied sample sizes from 100 to 1000. The detailed simulation scenarios are outlined in Supplementary Table S2.

Next, we simulated a scenario without DA taxa, where the observed difference in counts was completely due to sequencing depth. For this scenario, we generated the sequencing depth of the two groups separately at different levels. The sequencing depth for samples in the second group was, on average, three times greater than for the first group ($X_j \sim \mathbf{U}(3a, 3b), j \in$ **group 2**). We forced all taxa in both groups to have identical model coefficients (i.e., $\beta_{i,2} = 0$).

We normalized the simulated raw counts with TaxaNorm and several other methods, namely ANCOM-BC, TSS, TMM, CSS, and Wrench (Supplementary Table S3). We conducted DA testing with the post-normalized counts for each taxon using the Wilcoxon rank-sum test, adjusted p-values using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) for false discovery rate (FDR) control at 0.05, and compared performance in terms of power and FDR for detecting true DA taxa. We repeated each simulation scenario procedure 1000 times.

4. Results

4.1. Simulation studies results

Performance for diagnosis tests. When assessing the performance of the model diagnosis tests proposed in Section 2.3, TaxaNorm showed adequate power and good control for type I error in both prevalence and equivalence tests. When no sequencing effect exists for any taxa, the prevalence test controlled type I error at a nominal level of 5% (Table 1). Power is lacking with smaller sample sizes when the majority of taxa are not influenced by sequencing depth. However, the prevalence test showed increased power for identifying sequencing depth with larger sample sizes. This is expected because TaxaNorm relies on the regression model, which requires a sufficient sample size for accurate parameter estimation. Specifically, with a sufficiently large sample size, power will reach 100% (Table 1). When the sequencing depth effect exists for at least one taxon but varies across taxa, the equivalence test is highly stable with power over 90% (Table 1). Under the null hypothesis, with a consistent sequencing depth effect for all taxa, the type I error rate for the equivalence test is also under 10% (Table 1). In addition, TaxaNorm continues to produce reliable parameter estimations for $\beta_{i,0}$ and $\beta_{i,1}$ with the EM algorithm (Supplementary Figure S6). Estimation for zero and dispersion parameters is not perfectly unbiased, but this is within expectations because these estimates are unstable for the ZINB model (Jiang et al., 2023). As shown in simulation results for the DA analysis outlined later, downstream analysis is not affected even though the parameters are biased.

Performance for DA analysis. The simulation results indicate that, in various settings, TaxaNorm has better overall performance for balancing power and FDR compared to existing methods. When biological differences are small (fold change from $U(2, 4)$), only TaxaNorm, ANCOM-BC, and CSS had a well-controlled FDR at a 5% nominal level regardless of the sample size and proportion of DA taxa (Figure 3b, Supplementary Figure S1b and S2b). In particular, among these three, TaxaNorm is the most powerful method when the proportion of DA taxa is smaller (Figure 3a). TaxaNorm becomes conservative with a smaller sample size, which is within expectations, but maintains good power compared to CSS and ANCOM-BC. CSS is powerful with large sample sizes but loses a substantial amount of power when sample sizes decrease (Supplementary Figure S1a and S2a). Only with 50% DA taxa does CSS have higher power than TaxaNorm. Although ANCOM-BC has the best performance for controlling FDR, its power is not as high as TaxaNorm for identifying true DA taxa. For other methods, TSS has similar power as TaxaNorm but is much more conservative with a large proportion of DA taxa. TMM has the highest power for most scenarios. Wrench and CSS have similar performance. However, all these methods have a severely inflated FDR with a range from 20% to 60% in all scenarios (Figure 3b, Supplementary Figure S1b - S5b). The FDR is uncontrolled even with larger samples size and fewer DA included. Thus, their performance is not optimal considering the balance between power and FDR control. Interestingly, with a higher percentage of DA taxa, most methods, including TaxaNorm, lose power. However, CSS and Wrench have increased power, indicating that they may be more robust for data with larger variation. For larger biological differences (fold change from $U(4, 10)$), TaxaNorm also has compelling performance compared to existing methods (Supplementary Figure S3, S4 and S5).

For datasets without DA taxa, we calculated the log fold-change values for each taxon. Ideally, in normalized data, the estimated log fold-change for post-normalized data is around 0 (Figure 4). As shown in Figure 4, compared to the other methods, TaxaNorm provides

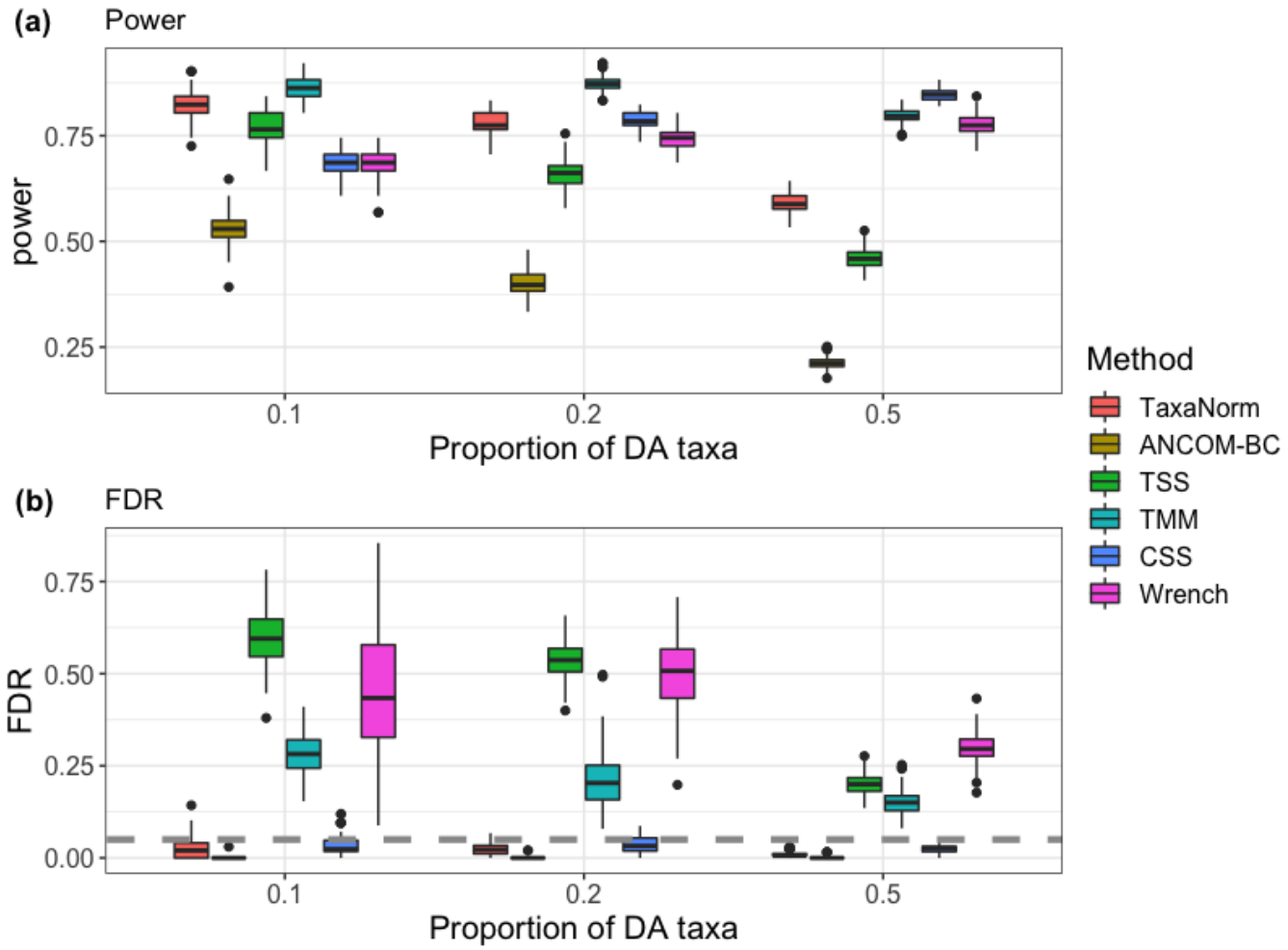

Fig. 3: Comparison of normalization methods in terms of **(a)** power and **(b)** FDR for identifying DA taxa in simulated datasets (sample size = 500 per group). The biological difference between groups is set to be small (fold-change from $U(2, 4)$). Each panel shows a simulation scenario with various percentages of DA taxa on the x-axis. The corresponding methods are denoted by colors as shown in the legend. The Benjamini-Hochberg procedure was used to adjust for the multiple testing burden with 5% as the nominal FDR level (dashed line).

unbiased estimates of log fold-change with the smallest variation compared to other methods. ANCOM-BC and CSS also yield unbiased estimates of log fold-changes, but the variations are larger, and TMM and Wrench yield biased estimates.

4.2. Real data application

4.2.1. Dataset

We applied our method to normalize raw taxa counts in data from the Human Microbiome Project (HMP) (The Human Microbiome Project Consortium, 2012a,b). The samples were from a human microbiome catalog comprising samples collected from five major body sites (oral cavity, nasal cavity, skin, gastrointestinal tract, and urogenital tract) in 300 healthy individuals aged between 18 and 40 years. Subjects provided samples at up to three visits, and taxonomic profiles were generated from 16S rRNA gene amplicon sequencing (16S) and whole-genome shotgun sequencing (WGS). Reads were deposited into the Data Analysis and Coordination Center, and the taxa counts and metadata can be downloaded from `https://www.hmpdacc.org/hmp/`. The data collection protocol and samples are described elsewhere (The Human Microbiome Project Consortium, 2012a,b).

4.2.2. Results

We first applied various normalization methods to data on the HMP gut mircobiota samples to determine their performance for eliminating the effect of sequencing depth. Figure 1 presents the relationship between taxa counts and sequencing depth before and after normalization with various methods, including TaxaNorm. Two taxa (*Dehalobacterium* and *Bacteroides*) with different abundance and sparsity levels were used for illustration. As expected, the raw taxa counts increased with sequencing depth, but the trends were not identical (Figure 1a).

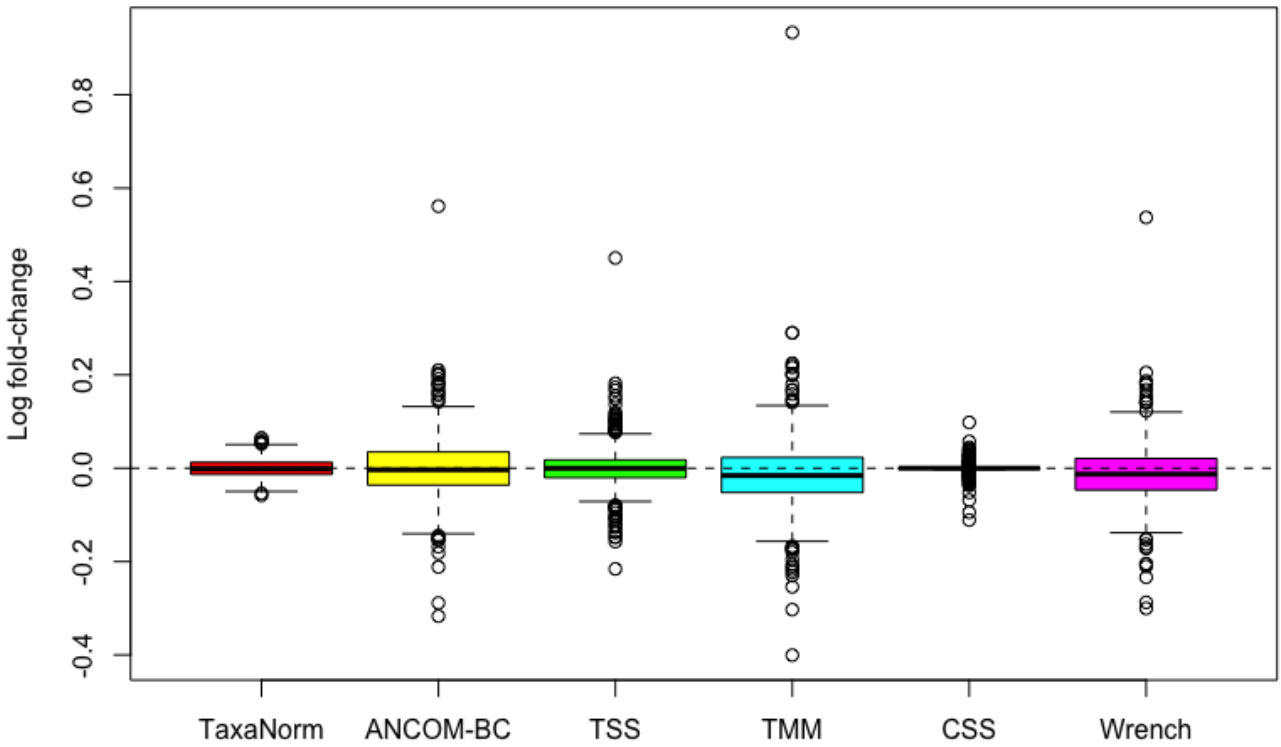

Fig. 4: Comparison of normalization methods in terms of estimated fold-changes in simulated datasets. No true DA taxa were simulated. The boxplots show the log fold-change values for all taxa for the two simulated groups after applying each normalization method. The dashed line denotes a log fold-change value of 0.

However, employing existing scaling methods with a global sample-specific size factor resulted in only partial removal of the sequencing depth effect, regardless of whether the methods were developed with RNA-seq data or microbiome data, while TaxaNorm completely removed the influence of sequencing depth on both taxa (Figure 1d). With the prevalence and equivalence tests proposed in Section 2.3, we found sequencing depth has a non-zero effect on at least one taxon (p-value $< 1e^{-10}$), and the effect was not identical across taxa (p-value $< 1e^{-10}$).

We then performed non-metric multidimensional scaling (NMDS) to assess how TaxaNorm and other normalization methods influence downstream analysis to distinguish samples by phenotype. For this, we restricted our analysis to the WGS data, which consists of 1,192 taxa obtained from 749 samples in five sites on the human body, including stool (n=147), skin (n=27), vagina (n=67), oral cavity (n=415), and nasal cavity (n=93). As shown in Figure 5, TaxaNorm showed a clear clustering pattern for samples from different body sites, particularly for skin and nasal cavity samples (Figure 5a) compared to other methods. For ANCOM-BC, TSS, CSS, and Wrench, the skin samples were mixed with nasal cavity samples (Figure 5b, 5c, 5e, 5f). CSS assigned a few skin samples apart from the others (Figure 5d, 5e). TSS, TMM, CSS, and Wrench had poor performance for differentiating vagina samples and oral cavity samples (Figure 5c-f). ANCOM-BC provided a poor classification of nasal cavity and oral cavity samples (Figure 5b). Additionally, TaxaNorm produced the largest between-group sum of squares (BSS) value while Wrench produced the smallest BSS value, indicating worse performance of Wrench for clustering.

5. Discussion

Normalizing microbiome data prior to downstream analysis is crucial because of the potential bias introduced by variations in sequencing depth across samples, which can result in undesirable and misleading conclusions regarding underlying biological mechanisms (McMurdie and Holmes, 2014; Weiss et al., 2017; Lin and Peddada, 2020a). Normalization is conducted to remove such systematic effects so that

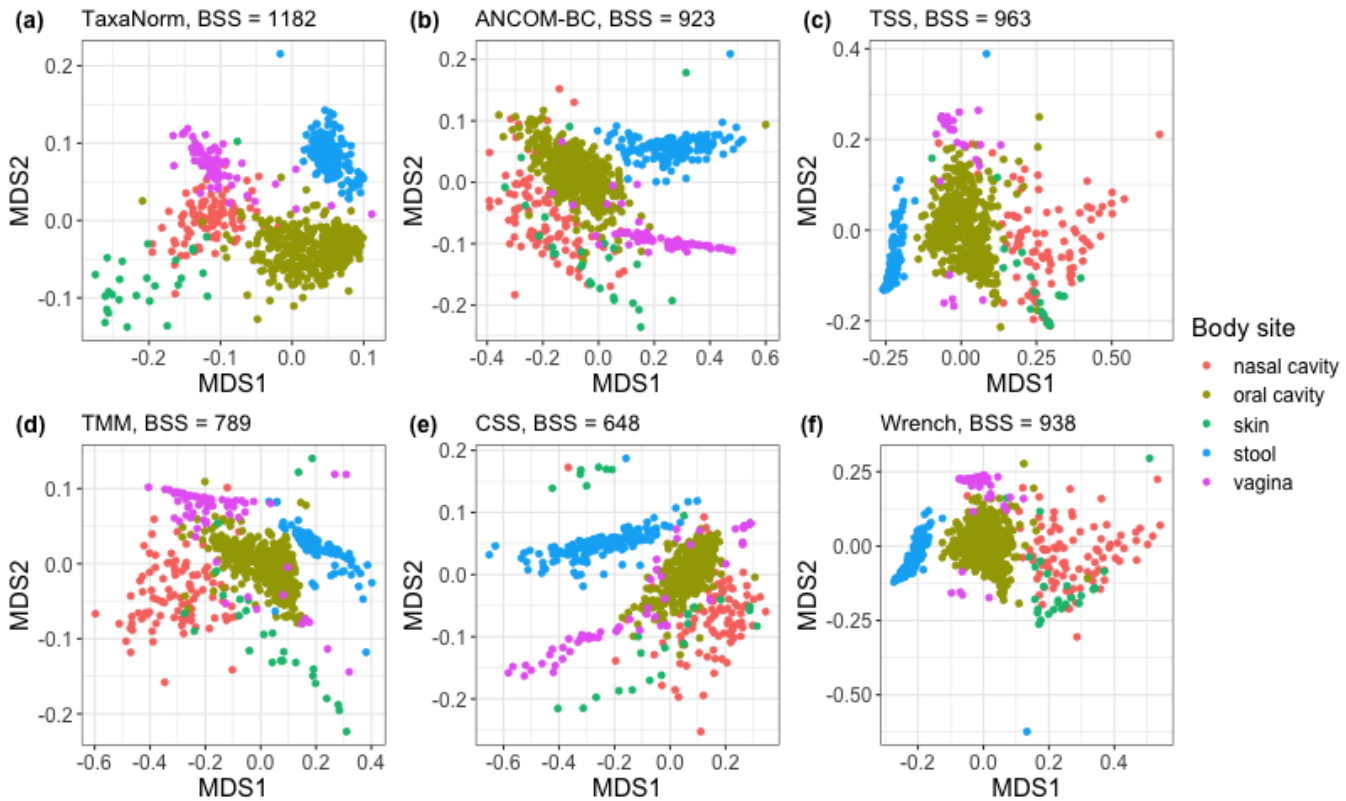

Fig. 5: NMDS visualizations for normalized HMP data. Two NMDS coordinates were used to evaluate the performance of various normalization methods: **(a)** TaxaNorm, **(b)** ANCOM-BC, **(c)** TSS, **(d)** TMM, **(e)** CSS, **(f)** Wrench. Sample type is denoted by colors according to the legend (nasal cavity, oral cavity, skin, stool, vagina). Values for between-group sum of squares (BSS) are also given.

all samples are on the same scale and independent of sequencing depth and thus the results will reflect true differences in underlying biology. However, existing normalization methods based on scaling do not sufficiently remove this effect because they violate the varying sequencing efficiency assumption and yield an elevated FDR, which results in loss of power in downstream analysis. Further, McMurdie and Holmes (2014) demonstrated that other non-parametric normalization methods such as rarefaction are inadmissible and result in loss of information due to over-dispersion of the taxa count, decreasing their power.

Our proposed TaxaNorm, a novel taxa-specific normalization method for microbiome data, addresses these drawbacks and has several advantages compared to existing methods, namely taking into account the varying effects of sequencing depth across taxa. Because of its demonstrated utility in advanced sequencing experiments, we explored ZINB regression, which models taxa count with sequencing depth as a covariate and uses the residuals as normalized count values. To better differentiate the structure of excess zeros and accommodate sample heterogeneity, we finalized our model in a flexible setting by jointly modeling zero counts and sample-specific dispersion. The results of simulation studies show that TaxaNorm has good performance for identifying true DA taxa with a low FDR. In real-data applications, TaxaNorm yielded a better grouping of samples from the same biological origin (i.e., same microbial community). These results indicate that TaxaNorm offers improved accuracy and efficiency for downstream analysis and visualization compared to existing normalization methods. In addition to correcting for bias due to uneven sequencing depth, we propose two powerful tests to rigorously check the goodness-of-fit for TaxaNorm. These tests are used for model diagnosis on a per-taxon basis, and simulation results show that TaxaNorm is robust even under a non-model assumption.

Importantly, TaxaNorm is not limited to microbiome data and can be applied in any omics data produced from sequencing technologies. This functionality is particularly important in light of the recent movement to collect genomics data in epidemiologic and environmental health sciences studies. Since TaxaNorm is built on a regression framework, its performance is affected by sample size and outliers. For the best performance, we recommend applying our method in data with a moderate sample size and conducting winsorization for any extreme taxa counts. Considering Bayesian regression with a prior when estimating the parameters will improve the model fit in the future. In future work, TaxaNorm can be also extended by including batch effects or other taxa-dependent covariates such as GC-content and genome length, which has been shown to affect taxa abundances from sequencing (McLaren et al., 2019; Sims et al., 2014; Browne et al., 2020; Ross et al., 2013). Due to the flexible specification of our model setting, it is convenient and easy to include new covariates. Another potential extension is incorporating a phylogenetic tree to account for dependency between taxa. This would enable information on taxa with similar evolutionary paths to be pooled and parameters to be regularized, such that similar taxa would share the same regression coefficients in the TaxaNorm model. This would improve computational efficiency. Further, with the increasing popularity of longitudinal

studies, the extension of TaxaNorm to mixed effects modeling on ZINB regression would be useful. Finally, reclaiming the true absolute feature counts (e.g., taxa abundance for microbiome data or gene expression for RNA-seq data), regardless of sequencing depth, using an advanced normalization algorithm can enable scientists to avoid deep sequencing and thus reduce the high costs associated with the technique. Further, additional datasets could be explored to further expand and validate TaxaNorm and improve existing normalization methods.

Acknowledgements

The authors thank Frank Day and Deepak Mav of NIEHS for expert computational assistance. We also appreciate Shyamal D. Peddada and Thomas A. Randall for their generous suggestions on manuscript writing for NIH internal review.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS).

Data availability

The HMP data can be downloaded from `https://www.hmpdacc.org/hmp/`.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

ZW initiated the research question, formulated the model development, conducted data analysis, and drafted the manuscript. DL created the R package. SZ and AMR provided statistical input on model specification. All authors edited and approved the final manuscript.

Supplementary Data

Supplementary materials are available on the journal's website.

References

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2): 139–160, 1982.

W. Barcik, R. C. Boutin, M. Sokolowska, and B. B. Finlay. The role of lung and gut microbiota in the pathology of asthma. *Immunity*, 52(2):241–255, 2020.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

P. D. Browne, T. K. Nielsen, W. Kot, A. Aggerholm, M. T. P. Gilbert, L. Puetz, M. Rasmussen, A. Zervas, and L. H. Hansen. GC bias affects genomic and metagenomic reconstructions, underrepresenting gc-poor organisms. *GigaScience*, 9(2):giaa008, 2020.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11(1):1–13, 2010.

J. Chen, N. Chia, K. R. Kalari, J. Z. Yao, M. Novotna, M. M. Paz Soldan, D. H. Luckey, E. V. Marietta, P. R. Jeraldo, X. Chen, et al. Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Scientific Reports*, 6(1):28484, 2016.

I. Cho and M. J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

P. I. Costea, G. Zeller, S. Sunagawa, and P. Bork. A fair comparison. *Nature Methods*, 11(4):359–359, 2014.

J. M. Di Bella, Y. Bao, G. B. Gloor, J. P. Burton, and G. Reid. High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, 95(3):401–414, 2013.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

A. D. Fernandes, J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):1–13, 2014.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

J. M. Gonzalez, M. C. Portillo, P. Belda-Ferre, and A. Mira. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PloS One*, 7(1):e29973, 2012.

J. B. Hughes and J. J. Hellmann. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods in Enzymology*, 397:292–308, 2005.

R. Jiang, X. Zhan, and T. Wang. A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data. *Journal of the American Statistical Association*, 118(542):792–804, 2023.

J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, et al. Evaluation of 16s rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):1–11, 2019.

A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology*, 8:2114, 2017.

O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, 9 (1):e1002863, 2013.

M. S. Kumar, E. V. Slud, K. Okrah, S. C. Hicks, S. Hannenhalli, and H. Corrada Bravo. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics*, 19(1):1–23, 2018.

H. Lin and S. D. Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):1–11, 2020a.

H. Lin and S. D. Peddada. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms and Microbiomes*, 6(1):1–13, 2020b.

H. Lin, M. Eggesbø, and S. D. Peddada. Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. *Nature Communications*, 13(1):1–16, 2022.

Y.-X. Liu, Y. Qin, T. Chen, M. Lu, X. Qian, X. Guo, and Y. Bai. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell*, 12(5):315–330, 2021.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26(1):27663, 2015.

M. R. McLaren, A. D. Willis, and B. J. Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8: e46923, 2019.

P. J. McMurdie and S. Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10 (4):e1003531, 2014.

J. T. Morton, C. Marotz, A. Washburne, J. Silverman, L. S. Zaramela, A. Edlund, K. Zengler, and R. Knight. Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10(1):1–11, 2019.

J. A. Navas-Molina, J. M. Peralta-Sánchez, A. González, P. J. McMurdie, Y. Vázquez-Baeza, Z. Xu, L. K. Ursell, C. Lauber, H. Zhou, S. J. Song, et al. Advancing our understanding of the human microbiome using QIIME. In *Methods in Enzymology*, volume 531, pages 371–444. Elsevier, 2013.

J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013.

J. N. Paulson, H. C. Bravo, and M. Pop. Reply to:" a fair comparison". *Nature Methods*, 11(4):359–360, 2014.

J. Pereira-Marques, A. Hout, R. M. Ferreira, M. Weber, I. Pinto-Ribeiro, L.-J. Van Doorn, C. W. Knetsch, and C. Figueiredo. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology*, 10:1277, 2019.

C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844, 2017.

T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018.

R. Ranjan, A. Rani, A. Metwally, H. S. McGee, and D. L. Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4):967–977, 2016.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, 2010.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):1–20, 2013.

J. U. Scher, A. Sczesnak, R. S. Longman, N. Segata, C. Ubeda, C. Bielski, T. Rostron, V. Cerundolo, E. G. Pamer, S. B. Abramson, et al. Expansion of intestinal prevotella copri correlates with enhanced susceptibility to arthritis. *eLife*, 2:e01202, 2013.

J. D. Silverman, K. Roche, S. Mukherjee, and L. A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020.

D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486:215–221, 6 2012a. ISSN 00280836. doi: 10.1038/nature11209.

The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402): 207–214, 2012b.

F. v. Wintzingerode, U. B. Göbel, and E. Stackebrandt. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, 21(3):213–229, 1997.

S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):1–18, 2017.

J.-Y. Wu, X.-T. Jiang, Y.-X. Jiang, S.-Y. Lu, F. Zou, and H.-W. Zhou. Effects of polymerase, template dilution and cycle number on PCR based 16s rRNA diversity analysis using the deep sequencing method. *BMC Microbiology*, 10(1):1–7, 2010.

R. Zaheer, N. Noyes, R. Ortega Polo, S. R. Cook, E. Marinier, G. Van Domselaar, K. E. Belk, P. S. Morley, and T. A. McAllister. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, 8(1):1–11, 2018.