Genetics and population analysis

An omnibus test for differential distribution analysis of microbiome sequencing data

Jun Chen1,2,*,†, Emily King1,3,†, Rebecca Deek4,†, Zhi Wei5,*, Yue Yu1,6, Diane Grill1 and Karla Ballman7

1Division of Biomedical Statistics and Informatics, 2Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA, 3Department of Statistics, Iowa State University, Ames, IA 50011, USA, 4Department of Biostatistics, Columbia University, New York, NY 10032, USA, 5Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA, 6College of Medical Informatics, Chongqing Medical University, Chongqing, China and 7Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY 10065, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Abstract

Motivation: One objective of human microbiome studies is to identify differentially abundant microbes across biological conditions. Previous statistical methods focus on detecting the shift in the abundance and/or prevalence of the microbes and treat the dispersion (spread of the data) as a nuisance. These methods also assume that the dispersion is the same across conditions, an assumption which may not hold in presence of sample heterogeneity. Moreover, the widespread outliers in the microbiome sequencing data make existing parametric models not overly robust. Therefore, a robust and powerful method that allows covariate-dependent dispersion and addresses outliers is still needed for differential abundance analysis.

Results: We introduce a novel test for differential distribution analysis of microbiome sequencing data by jointly testing the abundance, prevalence and dispersion. The test is built on a zero-inflated negative binomial regression model and winsorized count data to account for zero-inflation and outliers. Using simulated data and real microbiome sequencing datasets, we show that our test is robust across various biological conditions and overall more powerful than previous methods.

Availability and implementation: R package is available at https://github.com/jchen1981/MicrobiomeDDA.

Contact: chen.jun2@mayo.edu or zhiwei@njit.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

A central goal of human microbiome studies is to identify microbial taxa associated with certain biological or clinical conditions such as disease state. The identified taxa can provide insights into disease etiology as well as have the potential for use as biomarkers for disease diagnosis and prevention, and if causal, as therapeutic targets (Virgin and Todd, 2011). The development of next generation sequencing technologies enables culture-independent investigations of the role of the human microbiome in health and disease via direct DNA sequencing (Cho and Blaser, 2012). Both marker-gene targeted sequencing and whole-genome sequencing have been used to study the human microbiome, producing an abundance table based on which differential abundance analysis is performed (Franzosa et al., 2015).

Various statistical methods have been developed to perform differential abundance analysis, ranging from simple adaptation of the t test (Metastats) (Paulson et al., 2011) to more sophisticated statistical tests based on zero-inflated Gaussian model (metagenomeSeq, RAIDa) and log ratio statistics (ANCOM) (Mandal et al., 2015; Paulson et al., 2013; Sohn et al., 2015). These methods have evolved
with regard to addressing the special characteristics of the microbiome sequencing data, which are zero-inflated count data with variable library sizes (sequencing depths). To address the variable library sizes, traditional methods involve either normalizing the count data into proportion data (divided by the total sum) or rarefying the data into equal library sizes. McMurdie and Holmes (2014) recently showed the inadmissibility of methods based on rarefied or proportion data and suggested directly applying the negative-binomial (NB) based methods, such as DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) for RNA-seq data, to the taxa count data. Compared to RNA-seq data, the microbiome data contain a large number of zeros (typically > 50%). Many attempts have been made to tackle the excessive zeros including the application of hurdle and zero-inflated models (Chen and Li, 2016; Paulson et al., 2013; Wagner et al., 2011; Zhang et al., 2016). A recent comprehensive evaluation of competing models for zero-inflated microbiome data has revealed that count-based hurdle and zero-inflated models have well controlled Type I errors, higher power, and better goodness of fit measures, and are more accurate and efficient in the parameter estimation (Xu et al., 2015). Compared to the hurdle model, the zero-inflated model, which assumes that the observed zeros are a mixture of ‘structural zeros’ (due to physical absence) and ‘sampling zeros’ (due to under-sampling), is biologically more interpretable. The model is also consistent with the observation that the human microbiome has a small ‘core’ as revealed by Human Microbiome Project (HMP) (Huse et al., 2012), suggesting that at least part of the observed zeros are due to physical absence and cannot be explained by under-sampling. The zero-inflated model also allows us to investigate not only the possible association of biological conditions with the abundance levels, but also their associations with the probability of structural zeros. Converging statistical and biological evidence suggests that the count-based zero-inflated model is more appropriate for microbiome data.

Previous zero-inflated models for differential abundance analysis focused on testing the change in the abundance (mean of the non-zero component) and/or prevalence (probability of the nonzero component), treating the dispersion as a nuisance parameter (Chen and Li, 2016; Fang et al., 2016; Zhang et al., 2016). These models also assume a common dispersion parameter and do not allow the dispersion to depend on covariates such as disease condition. However, the fixed dispersion assumption is very restrictive and differential dispersion has been frequently observed in microbiome studies (Chen et al., 2016; Scher et al., 2013). Increased dispersion can be explained by the existence of microbiome-environment interaction or sample heterogeneity. For instance, the abundance of Prevotella corporis has been shown to depend on the HLA genotype in new-onset rheumatoid arthritis (NORA) subjects, resulting in increased dispersion in the RA group (Scher et al., 2013) (Fig. 1). Increased dispersion has also been observed in patients with multiple sclerosis, which consist of two subgroups of active and remission disease states (Chen et al., 2016). Besides sample heterogeneity, increased dispersion could also be due to loss of regulation of the normal microbial community, a potential consequence of dysbiosis (Kelsen et al., 2012). Therefore, allowing covariate-dependent dispersion can improve the robustness of zero-inflated models.

We thus propose a robust and powerful framework of differential analysis of microbiome data based on a zero-inflated negative binomial (ZINB) regression model, where we allow all the parameters depend on covariates. We also propose an omnibus test of all the three parameters—prevalence, abundance and dispersion, to identify associated microbial taxa. Since these parameters characterize the entire count distribution, our method represents a general approach for ‘differential distribution analysis’, extending the previous frameworks which target for only differential abundance analysis. Including the dispersion in the test can help to detect microbial taxa of direct biological/ecological relevance, which has already been shown to be instrumental to finding altered gene expression network and genetic epistasis in human genomic studies (Brown et al., 2014; Ho et al., 2008). Furthermore, it can also increase the detection power in presence of sample heterogeneity, where both mean and dispersion are affected (Fig. 1). In an extreme case of the existence of subgroups of opposite effects, previous tests will become powerless without taking into account the dispersion. These differentially dispersed taxa can also have predictive power for practical use, where the deviation from the mean could be the predictor.

Besides zero-inflation, the microbiome data are riddled with outliers (overly abundant taxa), leading to inflated Type I error and incorrect assessment of the statistical significance for many parametric models. Unfortunately, existing methods for differential abundance analysis for microbiome data have limited ability to deal with outliers. Though DESeq2 and the robust edgeR have proposed ways to deal with outliers, the effectiveness for microbiome data has not been assessed. In our differential distribution analysis framework, we thus also seek to find the most effective way to address the outliers to further improve the robustness of our method.

Using simulated data and real microbiome sequencing datasets, we show that our proposed framework is robust across various biological conditions and overall more powerful than previous methods.

### 2 Materials and methods

#### 2.1 Model setup

Let $c_{ij}$ be the number of reads from taxon $j$ in the $i$th sample, we model $c_{ij}$ based on the ZINB model

$$
\log(\mathbb{E}(c_{ij}|\mu_i, \phi_i)) = \mu_i + \phi_i \cdot \log(\mu_i) + (1 - \phi_i) \cdot \log(c_{ij} + \phi_i).
$$

Fig. 1. An example of differential dispersion. The dispersion of the abundance of Prevotella is different between healthy controls (HLT) and new-onset rheumatoid arthritis (NORA) patients. The increased dispersion of NORA group can be explained by sample heterogeneity. Solid box: Prevotella abundance distribution of HLT versus NORA patients. Dashed box: NORA group contains two subgroups stratified based on their genotypes (SE). The data was taken from Scher et al. (2013)

$$
\mathbb{E}(c_{ij}|\mu_i, \phi_i) = \mu_i + \phi_i \cdot \log(\mu_i).
$$

$$
\log(\mathbb{E}(c_{ij}|\mu_i, \phi_i)) = \mu_i + \phi_i \cdot \log(\mu_i) + (1 - \phi_i) \cdot \log(c_{ij} + \phi_i).
$$

$$
\begin{align*}
\log(\mathbb{E}(c_{ij}|\mu_i, \phi_i)) &= \mu_i + \phi_i \cdot \log(\mu_i) + (1 - \phi_i) \cdot \log(c_{ij} + \phi_i), \\
\log(\mathbb{E}(c_{ij}|\mu_i, \phi_i)) &= \mu_i + \phi_i \cdot \log(\mu_i).
\end{align*}
$$
which is a mixture of a point mass at zero and a negative binomial (NB) distribution of the form

\[ f_{nb}(c_j|\mu_j, \phi_j) = \frac{\Gamma(c_j + \frac{1}{\phi_j})}{\Gamma\left(c_j + 1\right) \Gamma\left(\frac{1}{\phi_j}\right)} \left(1 + \frac{1}{\phi_j} \mu_j\right)^{-c_j} \left(1 + \frac{1}{\phi_j} \mu_j\right) \frac{1}{c_j!} \]

(2)

The ZINB model is completely specified by the prevalence, abundance and dispersion parameters: \( \mu_j \) and \( \phi_j \), which characterize the probability of structural zeros (excess zero probability), and the mean and dispersion of the NB distribution, respectively.

Assume we also have covariates \( x_i \) for \( r \)th sample. We employ a generalized linear model to connect the parameters with the covariates using the following link functions:

\[
\begin{align*}
\logit(p_{ij}) &= x_{ij}^T \beta_j \\
\log(\pi_{ij}) &= x_{ij}^T \beta_j \\
\log(\phi_j) &= x_{ij}^T \gamma_j,
\end{align*}
\]

where \( \pi_{ij} = \mu_{ij}/s_i \), \( s_i \), \( \beta_j \), and \( \gamma_j \) determine the baseline properties for taxon \( j \), and \( x_j \), \( \beta_j \), and \( \gamma_j \) are regression coefficients for the prevalence, abundance and dispersion parameters, respectively. The parameter \( s_i \) is the scale factor, also known as size factor or normalization factor. It quantifies the library size (sequencing depth) of the sample, and is used to address the variable library sizes across samples. The above formulation is equivalent to including a log offset in the log link function for \( \mu_{ij} \). Though we use the same covariates \( x_i \) for simplicity of notation, in principle, covariates could be different for the three parameters. For simple two-group comparison, \( x_i \) represents the group indicator with \( x_i = 0 \) if the sample comes from the reference group, and \( x_i = 1 \) otherwise.

2.2 Model fitting

Denote all the parameters to be estimated as \( \theta = (s, \mu_j, \beta_j, \gamma_j)^T \). Then given the observed data \( (c_{ij}, i = 1, \ldots, n) \), the likelihood function of \( \theta \) is given by

\[
L_n(\theta) = \prod_{i=1}^{n} p_{ij}(\theta) I_0(c_{ij}) + (1 - p_{ij}(\theta)) f_{nb}(c_{ij}|\mu_{ij}(\theta), \phi_{ij}(\theta))
\]

We can maximize the log-likelihood function \( L_n(\theta) \) to obtain the maximum likelihood estimation (MLE) of \( \theta \). The direct maximizer of \( L_n(\theta) \) is numerically not stable. We turn to using an EM algorithm to obtain the ML estimates.

We introduce the latent variable \( z_{ij} \), which describes whether \( c_{ij} \) is from the NB distribution or the zero point mass. Let \( z_{ij} = 1 \) if \( c_{ij} \) is from the zero point mass, \( z_{ij} = 0 \) if it is from the NB distribution. Here we treat \( z_{ij} \) as the missing variable. The complete likelihood function under the augmented data, which include the original data \( c_{ij}s \) and the missing data \( z_{ij}s \), becomes

\[
L_n(\theta) = \prod_{i=1}^{n} \left[ p_{ij}(\theta)(1 - p_{ij}(\theta))^{1-z_{ij}} f_{nb}^{1-z_{ij}}(c_{ij}|\mu_{ij}(\theta), \phi_{ij}(\theta)) \right]
\]

and the log-likelihood function is

\[
L_n(\theta) = \sum_{i=1}^{n} \left[ z_{ij} \log p_{ij}(\theta) + (1 - z_{ij}) \log (1 - p_{ij}(\theta)) \right]
\]

\[ +(1 - z_{ij}) \log f_{nb}(c_{ij}|\mu_{ij}(\theta), \phi_{ij}(\theta)) \]

(3)

Using the spirit of EM algorithm by computing the conditional expectation (E-Step) and maximizing (M-Step) it over the parameter space, we have the following estimation procedure:

**E-Step.** After given an initial estimate \( \theta(0) \), calculate the conditional expectation \( Q(\theta, \theta(0)) = E[L_n(\theta|z_{ij}s, \theta(0))] \). With some algebra, we obtain

\[
Q(\theta, \theta(0)) = \sum_{i=1}^{n} \left[ a_i^{(0)} \log p_{ij}(\theta) + (1 - a_i^{(0)}) \log (1 - p_{ij}(\theta)) \right]
\]

\[ +(1 - a_i^{(0)}) \log f_{nb}(c_{ij}|\mu_{ij}(\theta), \phi_{ij}(\theta)) \]

(4)

where

\[
a_i^{(0)} = P_i(z_{ij} = 1|c_{ij}, \theta(0)) = \frac{p_{ij}(\theta(0)) I_0(c_{ij})}{p_{ij}(\theta(0)) I_0(c_{ij}) + (1 - p_{ij}(\theta(0))) \log f_{nb}(c_{ij}|\mu_{ij}(\theta(0)), \phi_{ij}(\theta(0)))}
\]

**M-Step.** Update the estimate \( \theta \) as \( \theta(1) \) as the solution to the score equation

\[
\frac{\partial}{\partial \theta} Q(\theta, \theta(0)) = 0.
\]

Then we can repeat the E-step and M-step for \( t = 1, 2, \ldots \), until the desired convergence is achieved. We can use generic optimization methods such as Newton-Raphson algorithm to obtain the M-step solution.

2.3 Omnibus test

The general hypothesis we want to test is

\[
\theta^A = \beta_j^A = \gamma_j^A = 0,
\]

where \( A \) indicates the coefficients for the covariates of interest. We use likelihood ratio (LR) test to conduct the proposed omnibus test, where we fit the model using MLE under both the null (reduced) and alternative (full) model. The LR statistic is asymptotically \( \chi^2 \) distributed with a degree of freedom (d.f.) being the difference of d.f. between the full and reduced model. For a two-group comparison, the omnibus test is reduced to test for whether a taxon differs in any of abundance, mean, or prevalence parameter between the two groups, that is, we are testing for

\[
\pi_1 = \pi_2, \beta_1 = \beta_2, \gamma_1 = \gamma_2.
\]

The LR statistic can be easily obtained by fitting the ZINB model to all samples, and the samples from each group separately. The LR test statistic has a \( \chi^2 \) distribution in this special case.

2.4 Scale factors

In the proposed ZINB regression model, we include the scale factor \( s_i \) to account for variable library sizes across samples. Calculating the scale factor has also been referred as ‘normalization’. The naive method is to use the total count as the scale factor (Total Sum Scaling, TSS). This strategy is still widely employed in the analysis of microbiome-seq data. However, in cases where some dominant taxa are differentially abundant, TSS will create strong ‘compositional’ effects, i.e. these non-differentially abundant taxa will appear to be differentially abundant with their relative abundances being driven up/down by these truly differentially abundant taxa. Therefore, it is desirable to derive a robust scale factor to capture the invariant (non-differential) part of the counts. A robust normalization procedure will reduce the false positive rates due to compositional effects. We will use Geometric Mean of Pairwise Ratios or GMPR, a new method also developed by our group, to calculate the scale factor, taking into account the uncertain source of the zeros (Chen and Chen, 2017). Comprehensive simulations and real data evaluations.
demonstrated the more robust performance of GMPR over previous approaches including the CSS (cumulative sum scaling), TMM and RLE with pseudo-counts added (Anders and Huber, 2010; Paulson et al., 2013; Robinson et al., 2010). Denote $c_{ij}$ as the count of the $j$th taxon in the $i$th sample. GMPR scale factor $s_i$ for sample $i$ is calculated as:

$$s_i = \left( \prod_{j=1}^{n} \text{Median}_{i \neq k} \frac{c_{ik}}{c_{jk}} \right)^{1/n}.$$

The basic idea of GMPR is that we conduct the pair-wise comparison first without the need for calculating the geometric mean for each taxon as implemented in RLE. Since only a small number of taxa (or none) are shared across samples, RLE is not stable or possible. For pair-wise comparison, we only focus on these taxa that are observed in both samples to have a reliable inference of the abundance ratio between samples. We then synthesize the pairwise ratios using geometric mean to obtain the final scale factor. To be noted, GMPR is a general method, which can be applied to any type of sequencing data.

2.5 Outliers

Although the ZINB model can capture large amounts of variation in counts, parametric bootstrap tests, where we generate the counts using the fitted model based on real taxa counts, suggest some taxa have outliers that are not accounted for by the fitted model (data not shown). These outliers may lead to elevated Type I error and thus must be accounted for. One way is to remove the outliers before the analysis. However, this strategy could bias the data in the opposite direction. We thus turn to the outlier replacement strategy, and consider the following two outlier replacement strategies:

1. Cook’s D: Fit a ZINB model to the data for a given taxon and compute Cook's D for each observation (Cook and Weisberg, 1982). Observations with Cook’s D exceeding $D_{crit}$ are flagged as outliers. Scaled counts are computed as $c_{ij}/s_i$. The outlier is replaced with the maximum non-outlying scaled count for its taxon rescaled by the scale factor for its sample.

2. Winsorization: Replace any observation exceeding the $z$ quantile of scaled counts $c_{ij}/s_i$ for a given taxon with the $z$ quantile of scaled counts for that taxon, rescaled by the scale factor.

3 Simulation studies

The omnibus test was applied to two classes of simulated datasets. The first set of simulation data consisted of count data from a single taxon at various levels of prevalence, abundance and dispersion and a wide range of change models were investigated. In this set of simulations, the library size/scale factor was assumed to be known, and the performance evaluation focused on Type I error control and power. The second set used modifications of a real gut microbiome dataset to assess performance on a more realistic multitaxon dataset, with or without outliers. Using a multitaxon dataset allows for assessment of compositional effects and FDR control, and comparison with more existing methods.

3.1 Single-taxon simulations

3.1.1 Simulation setup

All datasets were simulated from a ZINB model as given in Equation (1). Specifically, we considered 27 different sets of parameter values given in Table 1. These parameter values were combinations of low, moderate and high values for each of the three ZINB model parameters. Parameter choices were guided by the distribution of ZINB model parameters fitted to the 397 most prevalent OTUs (Operational Taxonomic Units) from a gut microbiome dataset of the COMBO study, which investigated the association between diet and the gut microbiome (Wu et al., 2011). The library sizes for samples were also drawn from those of the COMBO dataset, and ranged from 674 to 16 000. All datasets consisted of 100 individuals in two equally sized groups. There were 1000 datasets per simulation condition.

Our omnibus test was compared to four other competing methods in power and Type I error evaluations. These methods include (i) Fisher’s exact test, which tests differences in the proportion of zero counts between groups, (ii) the non-parametric Wilcoxon rank-sum test applied to counts normalized by library size, (iii) the twopart test (Wagner et al., 2011), which uses a test statistic of the sum of two other test statistics: a $Z$ statistic comparing the proportion of zeros in each group and a Wilcoxon rank-sum test statistic computed from nonzero counts and (iv) the negative binomial distribution-based likelihood ratio test (NB test) for difference in $\pi$ between groups. For each test, we set the nominal level of Type I error at 0.05 level.

3.1.2 Type I error control

We first simulated the null $H_0: \pi_1 = \pi_2, \rho_1 = \rho_2, \phi_1 = \phi_2$, to see if each test could control Type I error at the nominal level 0.05. Figure 2 shows the Type I error for each test. The omnibus test controlled the Type I error around the nominal level, ranging between 0.028 and 0.066. Type I error rates tended to be higher when the probability of the zero component was higher. When $\rho=0.01$, the test was slightly

### Table 1. Parameter values used in simulations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low</th>
<th>Intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$ (abundance parameter)</td>
<td>0.002</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>$\rho$ (prevalence parameter)</td>
<td>0.01</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>$\phi$ (dispersion parameter)</td>
<td>0.6</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>
conservative. Wilcoxon test controlled the Type I error very well while the Fisher’s exact test and the two-part test were conservative as expected. NB test was very conservative when the taxon’s prevalence was low, but the abundance was high once it was present. Under such scenarios, the NB model was inadequate to model the excessive zeros. We also simulated scenarios where there were systematic differences in sequencing depth between groups. Differential sequencing depths did not have a large influence on the Type I Error of the omnibus test and the NB test while the nonparametric tests had serious Type I error inflation under certain combinations of parameter values due to the discrete nature of count data and excessive zeros (Supplementary Fig. S1). This analysis further demonstrated the importance of modeling counts.

### 3.1.3 Power analysis

We next conducted a power analysis by simulating different alternatives. We fixed the parameter values in Group 1 and varied the parameter values in Group 2 with different effect sizes to create power curves. We investigated four scenarios, where an individual parameter (abundance, prevalence and dispersion) or all parameters were changed. Both low- and high-dispersion scenarios were studied (Group 1 for \( \phi = 0.6 \) versus 3).

Figure 3 shows the power of each test. Of the five tests, the omnibus test was the most robust: it was always the best or close to the best-performing test across the investigated scenarios. Other methods were not robust and performed poorly under certain scenarios. For example, Fisher’s exact test had no power to detect the change in the abundance parameter and NB test had limited power to detect change in either prevalence or dispersion. The power of NB test was also significantly worse than the omnibus test in detecting the abundance change under low dispersion setting. The performance of Wilcoxon rank sum test and two-part test were also poor under some settings (Fig. 3). In practice, we do not know the specific differential pattern in advance, and the omnibus test could identify a wide range of differential patterns, which could be otherwise missed by competing methods.

### 3.1.4 Additional simulations

We also compared our omnibus test to four other ZINB-based likelihood ratio tests, designed to test for change in specific components (1 d.f. LRTs for \( \pi, p \) and \( \phi \), and 2 d.f. LRT for both \( \pi \) and \( p_i \)), in order to assess the magnitude of power loss associated with the use of a three d.f. test when only a smaller number of parameters truly differed between groups. The basic observation was that the omnibus test was very robust, and the power loss was not substantial, compared to the best performing 1 d.f. or 2 d.f. LRTs targeted to specific components of the distribution (Fig. 4). We also studied the performance of the omnibus test under model over-specified, where we generated the data using NB distribution, and model misspecification, where we generated the counts using zero-inflated beta-binomial distribution. Under such settings, the omnibus test had once again been shown to be robust: it controlled the Type I error at the nominal level and the power loss was not significant, compared to the optimal test with the correct model assumption (Fig. 5).

### 3.2 Multiple-taxon simulation

#### 3.2.1 Simulation setup

Artificial datasets were generated based on the 397 most prevalent OTUs from the COMBO study (Wu et al., 2011). Specifically, the 397 OTUs in the dataset were present in at least 10% of samples and had a relative abundance of at least 0.002 in at least one sample.

ZINB models were fitted to the counts of these 397 OTUs to estimate the parameters. The parameter estimates were then used to generate artificial datasets consisting of counts of 397 OTUs in 98 individuals in two groups of size 49. To simulate differential OTUs, 5% of OTUs (with fitted \( p \leq 0.40, 0.0005 \leq \pi \leq 0.05, \phi \leq 10 \)) were randomly selected to have their counts in one group multiplied by a factor of \( k \) (\( k = 4 \)). The groups in which this occurred were randomly selected and potentially different for each differentially abundant OTU. The performance of the omnibus test was compared to four other methods designed for use on multiltaxon datasets: DESeq2, edgeR, metageneSeq and RAIDA (Love et al., 2014; Paulson et al., 2013; Sohn et al., 2015; Zhou et al., 2014). Their default setting and native normalization methods were used. For DESeq2, we did not allow outlier replacement since no outliers were simulated in the artificial datasets.
negatives, respectively. Then TPR, FPR and FDP were defined as number of true positives, true negatives, false positives and false
tive rates at different tistic curve, where true positive rates were plotted against false posi-
discovery proportion (FDP), and ROC (receiver operating character-
control, where FDR is defined as the expectation of false
ditions using false discovery rate control (FDR) procedure by

3.2.2 Performance measure
Each method output raw P-values of the tests for between-group dif-
ferences in each taxon. They were corrected for multiple compar-
sions using false discovery rate control (FDR) procedure by

3.2.3 Results
We first studied the ability of FDR control for the competing meth-
ods. As shown in Figure 6A, FDR was extremely elevated relative to
the nominal level using metageneomeSeq. FDR was also considerably
raised using edgeR. In addition, we tried the robust version of
edgeR and DESeq2. The true positive rate of the omnibus test was
higher than that of metageneomeSeq, edgeR and DESeq2 for a fixed
value of the false positive rate. The performance of RAIDA was
close to that of the omnibus test; however, the omnibus test was
more powerful at slightly high false positive rate level. Overall,
the omnibus test had the best FDR control ability and the best power.

We performed additional simulations, where we varied the
ZINB parameters instead of directly multiplying the counts
(Supplementary Fig. S2). In this simulation setup, the omnibus test
was far more powerful than the rest methods, partly due to the fact
that the data were generated based on our model assumptions.

To rule out the possibility that the omnibus test’s superior performance
was solely explained by GMPR normalization, we repeated the analysis using GMPR normalization for all methods that
used scale factors (DESeq2, edgeR and metageneomeSeq). Supplementary Figure S3 shows that using GMPR normalization
did improve the performance of DESeq2 and edgeR in terms of FDR control and power, while the performance of metageneomeSeq re-
mained similar. Nevertheless, the omnibus test was still more power-
ful than the competing methods, indicating that the power gains
could not be explained by using GMPR alone.

3.3 The effects of outliers
In the multiple-taxon simulation, the artificial datasets were gener-
ated based on ZINB model and no outliers were simulated. In real
datasets, many highly abundant outlier taxa are expected. Parametric models are easily influenced by outliers and, without
proper handling of outliers, many false associations may be pro-
duced. It is imperative to assess the effects of outliers on parametric
models and determine the best strategy to minimize their adverse ef-
facts before applying to real data. To study the outlier effects, indi-
viduals from the COMBO dataset were randomly assigned to either
Group 1 or Group 2, followed by differential abundance analysis on
the 397 most prevalent OTUs. This was intended to create random-
ized datasets in which there was no difference in count distributions
between groups for any taxon. Because there were no true between-
group differences in these simulations, the false positive rate was
simply the proportion of taxa found to differ significantly between
groups. If a Type I error of 5% was used to identify differential
taxa, we expected to see a false positive rate of 5% on average if a method was robust.

Figure 7 shows the false positive rates of each method using different strategies to address the outliers. When no outlier correction was performed, the average FPR of omnibus test was 0.107 at a nominal 0.05 level. Replacing observations with Cook’s D greater than 1 reduced its FPR to 0.077 and Winsorizing at the 0.97 quantile further reduced the FPR to 0.053. The Cook’s D cutoff was set to be 1 as suggested by Cook and Weisberg (1982). Without outlier replacement, DESeq2 and edgeR had FPR level considerably above the nominal level. Use of Cook’s D-based outlier replacement decreased FPR, though it remained above the nominal level. Use of winsorization further reduced FPR of these methods close to the nominal level. RAIDA (Sohn et al., 2015) is effective at controlling FPR close to the nominal level of 0.05 regardless of the outlier replacement method used, confirming the robustness of the method. All nonparametric tests controlled FPR near or below nominal levels as expected. metagenomeSeq had very high FPR compared to other methods, postulating FPR near or below nominal levels as expected.

thus compared our omnibus test to these two competing methods. We also compared to the Wilcoxon rank-sum test since it has been widely used in the microbiome literature. In order for the comparison to be fair, the data were winsorized at the 0.97 quantile before applying the methods. DESeq2’s default method for outlier replacement based on Cook’s D is disabled. The Wilcoxon rank-sum test was applied to rarefied count data.

4 Real data applications

4.1 Datasets

Finally, we applied our method to three real datasets from the study of the gut microbiota in a general population (COMBO), inflammatory bowel disease (IBD) and rheumatoid arthritis (RA) (Morgan et al., 2012; Scher et al., 2013; Wu et al., 2011). The COMBO and IBD datasets were downloaded from Qita (https://qiita.ucsd.edu/) with study ID 1011 and 1460, and the RA dataset was provided by the author. The OTU abundances were compared between normal weight (BMI < 25, n = 59) and overweight/obese subjects (BMI ≥ 25, n = 40) for COMBO data, between healthy (n = 18) and Crohn’s disease subjects (n = 62) for the IBD dataset, and between healthy (n = 21) and new-onset rheumatoid arthritis subjects (n = 44) for the RA dataset. For all the three datasets, only stool samples were included for the analysis. Any sample with less than 100 total reads was removed from the dataset. OTUs with prevalence less than 10% were excluded for testing. The dimensions of the original and processed datasets are given in Table 2.

4.2 Methods

Based on the simulation results, DESeq2 and RAIDA are overall more robust and powerful than edgeR and metagenomeSeq. We

Figure 7 shows the false positive rates of each method using different strategies to address the outliers. When no outlier correction was performed, the average FPR of omnibus test was 0.107 at a nominal 0.05 level. Replacing observations with Cook’s D greater than 1 reduced its FPR to 0.077 and Winsorizing at the 0.97 quantile further reduced the FPR to 0.053. The Cook’s D cutoff was set to be 1 as suggested by Cook and Weisberg (1982). Without outlier replacement, DESeq2 and edgeR had FPR level considerably above the nominal level. Use of Cook’s D-based outlier replacement decreased FPR, though it remained above the nominal level. Use of winsorization decreased FPR of these methods close to the nominal level. RAIDA (Sohn et al., 2015) is effective at controlling FPR close to the nominal level of 0.05 regardless of the outlier replacement method used, confirming the robustness of the method. All nonparametric tests controlled FPR near or below nominal levels as expected. metagenomeSeq had very high FPR compared to other methods, postulating FPR near or below nominal levels as expected.

thus compared our omnibus test to these two competing methods. We also compared to the Wilcoxon rank-sum test since it has been widely used in the microbiome literature. In order for the comparison to be fair, the data were winsorized at the 0.97 quantile before applying the methods. DESeq2’s default method for outlier replacement based on Cook’s D is disabled. The Wilcoxon rank-sum test was applied to rarefied count data.

4.3 Results

The numbers of significant associations found by each method analyzed at the OTU level with a 10% FDR cutoff, and their overlap were presented in Figure 8.

**Table 2. Number of samples and OTUs in each dataset before and after filtering**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Phenotype</th>
<th>#OTU_U</th>
<th>#OTU_F</th>
<th>#Sample_U</th>
<th>#Sample_F</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMBO</td>
<td>Overweight</td>
<td>1873</td>
<td>625</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>IBD</td>
<td>Crohn’s disease</td>
<td>2053</td>
<td>615</td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td>RA</td>
<td>NORA</td>
<td>954</td>
<td>438</td>
<td>72</td>
<td>72</td>
</tr>
</tbody>
</table>

U, unfiltered data; F, filtered data; NORA, new-onset rheumatoid arthritis.
found the largest number of significantly associated OTU (n = 35), followed by Wilcoxon test and DESeq2 (n = 33 for both). RAIDA found the smallest number of significant OTUs (n = 11). Interestingly, the omnibus test was the only method that could identify all the four OTUs enriched in NORA as reported by Scher et al. (2013) (Supplementary Table S3).

We also studied the overlap of the taxonomies for the identified associations (Supplementary Tables S1–S3). We could see a significant overlap between the methods since distinct OTUs may have the same high-level taxonomy. We note that the OTU/species level analysis might be biologically more meaningful than higher taxonomic levels since species from the same genus might have very different biological behaviors. We thus focused on the evaluation on the OTU level.

Though our major focus was the comparison of count-based parametric methods, it was also interesting to compare the omnibus test to LEfSe (Segata et al., 2011) and MaAsLin (Morgan et al., 2012), which were used in the original papers. Supplementary Figure S4 shows the overlap between the methods. LEfSe recovered the largest numbers of associations while MaAsLin recovered more associations than omnibus test on the IBD dataset and fewer associations on the COMBO dataset. For the RA dataset, omnibus test and MaAsLin performed similarly. We note that LEfSe, which also used Wilcoxon test, did not perform multiple testing correction while MaAsLin worked with proportion data, which could be sensitive to compositional effects especially in datasets such as the IBD dataset with large numbers of differential taxa.

5 Discussion
Differential abundance analysis of microbiome data is instrumental in identifying potential microbiome biomarkers for further biological validation. Nonparametric methods, such as Wilcoxon rank sum, have been widely used in the microbiome literature due to the robustness of their nonparametric assumption. However, there are many limitations of the nonparametric methods including inability to adjust covariates, lack of interpretable effect size definitions, and failure to use the full information. To account for differential library sizes in nonparametric methods, normalizing the count data into proportions or rarefying the data to the same library size has been suggested. McMurdie and Holmes (2014) demonstrate that these approaches are not admissible, and that count-based models may be more powerful. However, modeling the counts in the presence of excessive zeros and outliers is challenging, and many methods are not sufficiently robust (Mandal et al., 2015). To increase the robustness of the count-based approach, we propose a ZINB-based generalized linear regression framework that allows all the parameters (prevalence, abundance and dispersion) to depend on covariates, significantly extending the modeling capability of previous methods. An omnibus test that jointly tests all the three parameters is devised to detect differential taxa under a wide range of biological conditions. In real applications, we generally have little knowledge of the specific differential pattern. Therefore, our proposed omnibus test can potentially reveal differential taxa that are otherwise missed by other approaches. We also implemented an outlier replacement strategy to further improve the robustness of our framework. Our model is also very robust to model over-specification and misspecification as demonstrated by simulations. To be noted, the robustness of our approach is not at the cost of significant loss of power. Our method is overall more powerful than existing methods.

In our framework, we have evaluated two approaches for outlier replacement based on Cook’s distance and winsorization. We show that winsorization with the 97% quantile works well. This quantile hyper-parameter may be optimized for further improved performance. For example, the optimal quantile can be determined based on the false positive rate control on randomly shuffled datasets. Another possible approach for addressing outliers is through data reweighting: assign a weight to each data point based on its ‘outliness’ (Huber, 2011). Weighting can be done automatically and iteratively. In the first iteration, data points are assigned equal weights and model coefficients are estimated using an EM algorithm. At subsequent iterations, weights are recomputed so that points that are further from model predictions in the previous iteration are given lower weights. The process continues until the values of the coefficient estimates converge within a specified tolerance. Comparison of this approach to the proposed winsorization method will be a future research topic.

The proposed method depends on individual testing of all taxa without taking into account the phylogenetic tree information. The phylogenetic tree relates all the microbial taxa and provides important prior knowledge as to how these taxa are evolutionarily related: closely related taxa usually have similar biological characteristics and have a tendency to be co-differentiated. Thus, the phylogenetic tree information could be exploited, in principle, to further improve the detection power (Chen et al., 2013). In this regard, an empirical Bayes method is a promising approach to incorporate the tree information into testing (Smyth, 2004). Another alternative is to integrate the phylogenetic tree information into the multiple testing correction stage (Xiao et al., 2017).

Another challenge associated with differential abundance analysis of microbiome data is the compositional effects since the counts are constrained by the sequencing depth (Tsilimigras and Fodor, 2016). The compositional effects are more severe with the existence of a few highly abundant taxa. If the counts are normalized into proportions, the increase in the abundance of one dominant taxon will lead to the apparent decrease of the rest taxa. To address the compositional effects, three approaches are possible. The first is through an appropriate normalization method, such as the GMPR method used in the omnibus test, to produce a robust estimate of the library size (capturing the non-differential part) that is resistant to outliers and differential taxa. This strategy works best when the percentage of differential taxa is not large or the change is balanced between the groups. The second approach is to use a common and abundant non-differential taxon as a reference as implemented in the RAIDA method (Sohn et al., 2015). The difficulty is to identify such reference taxon and the variability of its abundance. The third approach is through the analysis of the patterns of pairwise log ratios (Mandal et al., 2015). For a truly differential taxon, its log ratio to all other taxa will vary with the variable of interest while the log ratio for the non-differential taxon will be a mixture of constant and varying components. The difficulty of this strategy is to accommodate excessive zeros and covariates as well as the power to separate these two distinct patterns under moderate effect sizes. In the proposed method, we used the first approach to address compositionality. We note that this approach may not be adequate when there are a large number of differential taxa. In such situation, a possible solution is to re-estimate GMPR size factor using the non-significant taxa. However, a principled approach to addressing compositional effects needs more investigation.

Both single- and multiple-taxon simulations demonstrated a generally good false positive control for the proposed method, compared to other parametric methods. However, the false positive control is not perfect: a trend of slightly inflated type I error has been observed for low-prevalence taxa in single-taxon simulation.
Differential distribution analysis of microbiome data

(Fig. 2). The slight type I error inflation translates into a slight FDR inflation for multiple-taxon simulation. We found that the cause of type I error inflation is due to the inaccurate approximation of the LRT-based asymptotic distribution for these rare taxa under a small sample size (Chi-squared distribution). To reduce the influence of these rare taxa, we recommend excluding these rare and low-abundance taxa from testing (e.g. only include taxa with more than 10 reads in at least 10% samples), and paying special attention to the small \(P\)-values for these rare taxa. These small \(P\)-values may be verified by a permutation test. Precise control of the type I error rate for these rare taxa warrants further study.

Our framework is very general and could be used to test different types of hypotheses. For example, one interesting test is to jointly test prevalence and abundance while allowing the dispersion to depend on covariates such as sequencing batches. As the microbiome research moves into more mechanistic studies, longitudinal and family-based microbiome data has been increasingly popular. A natural extension of our framework is to account for sample correlations in differential distribution analysis. Generalized mixed effects model (GLMM) can be used to this end (Breslow and Clayton, 1993; Zhang et al., 2017). However, addressing the zero-inflation in GLMM is not trivial and requires new methodology development.

Funding

This work was supported by Gerstner Family Career Development Award and Center for Individualized Medicine at Mayo Clinic.

Conflict of Interest: none declared.

References