RESEARCH

Distribution-based comprehensive evaluation of methods for differential expression analysis in metatranscriptomics

Hunyong Cho[1], Chuwen Liu[1], Boyang Tang[2], Bridget M. Lin[1], Jeffrey Roach[3], Apoena de Aguiar Ribeiro[4], Michael I. Love[1,5], Kimon Divaris[6,7] and Di Wu[1,8,9*]

[*]Correspondence: did@email.unc.edu
[1]Department of Biostatistics, University of North Carolina, Chapel Hill, US
Full list of author information is available at the end of the article

Abstract

Background: Measuring and understanding the function of the human microbiome is key for several aspects of health; however, the development of statistical methods specifically for the analysis of microbial gene expression (i.e., metatranscriptomics) is in its infancy. Many currently employed differential expression analysis methods have been designed for different data types and have not been evaluated in metatranscriptomics settings. To address this knowledge gap, we undertook a comprehensive evaluation and benchmarking of eight differential analysis methods for metatranscriptomics data.

Results: We used a combination of real and simulated metatranscriptomics data to evaluate the performance (i.e., model fit, Type-I error, and statistical power) of eight methods: log-normal (LN), logistic-beta (LB), MAST, Kruskal-Wallis, two-part Kruskal-Wallis, DESeq2, ANCOM-BC, and metagenomeSeq. The simulation was informed by supragingival biofilm microbiome data from 300 preschool-age children enrolled in a study of early childhood caries (ECC), whereas validations were sought in two additional datasets, including an ECC and an inflammatory bowel disease one. The LB test showed the highest power in both small and large sample sizes and reasonably controlled Type-I error. Contrarily, MAST was hampered by inflated Type-I error. Using LN and LB tests, we found that genes C8PHV7 and C8PEV7, harbored by the lactate-producing Campylobacter gracilis, had the strongest association with ECC.

Conclusion: This comprehensive model evaluation findings offer practical guidance for the selection of appropriate methods for rigorous analyses of differential expression in metatranscriptomics data. Selection of an optimal method is likely to increase the possibility of detecting true signals while minimizing the chance of claiming false ones.

Keywords: metatranscriptomics; metagenomics; differential expression; logistic-beta; log-normal; early childhood caries

1 Introduction

1.1 Significance

The human microbiome has emerged as an undeniable cornerstone for a multitude of health and disease outcomes. Important new insights have been recently gained regarding the pivotal role of microbial dysbiosis in conditions such as obesity, gut disease, cancer, oral and dental diseases [1, 2, 3, 4]. Contemporary investigations now seek to understand not only the composition of microbial communities (i.e., taxonomy) but also their functional activity. Taxonomy is typically ascertained by 16S rRNA or whole genome shotgun (WGS) sequencing with the latter offering several advantages over 16S sequencing, including a better phylogenetic resolution and information on genomic content (i.e., metagenomics). Microbial functional activity can be measured via RNAseq, i.e., metatranscriptomics [5] and metabolomics. Microbial gene expression and metabolism are where the "rubber meets the road", as they represent the viable and active members of the microbial community and reflect the biology underlying the microbiome's interactions with the host and the environment.

Gastrointestinal and oral health, above and beyond their common anatomical, functional, and biological similarities, are now both better understood using models of microbial symbiosis and dysbiosis, while microbiome links between the two have begun to emerge [6]. In both of these health and research fields, investigations involving metatrascriptomics have provided novel insights. An excellent example includes a multiomics study of inflammatory bowel disease (IBD) highlighting metatrascriptomics associations with the microbial community's temporal variability, taxonomic, and biochemical shifts [7]. In the oral health domain, Peterson et al. used metatranscriptomics to identify dominant functions associated with dental caries within the dental biofilm of 19 twin pairs—these functions supported the microbial communities' biochemical activities related to sugar metabolism and resistance to acid and oxidative stress [8]. In another study, metatranscriptomics was used to identify differences in the activity of the subgingival biofilm microbiome between 7 individuals with periodontitis and periodontally-health controls [9]. A notable recent review of metatranscriptomics analysis of the oral microbiome was recently reported by Duran-Pinedo [10].

Despite the increasing significance and availability of metatranscriptomics data, the development of tailored statistical analysis methods has not kept pace. There are few statistical analysis methods specifically designed to handle microbiome data, and a systematic evaluation of all existing methods that have been borrowed from other areas of high-throughput sequencing analysis (e.g., human studies) has yet to be undertaken.

1.2 State of existing differential microbial gene expression analysis methods

There are several special considerations in the analysis of microbiome data. These include but are not limited to data normalization [11], clustering of species or genes [12], alpha- and beta- diversity [13], differential abundance/expression (DA/DE) analysis [14, 15, 16], and metabolome-based pathway analysis [17]. Most methods that address these considerations have either been borrowed from analytic pipelines used in other high-throughput sequencing technologies—e.g., bulk RNAseq—or were developed for 16S or WGS metagenomics data. Most of these currently available methods to analyze metatranscriptome data rely on the joint mapping of microbial DNAseq and RNAseq data [17, 18, 19]. However, very few methods have been specifically developed for DE analysis of metatranscriptomics data [20, 21]. It follows that, while there is ample room for new metatranscriptomics analysis methods development, the existing approaches that have been developed for other data types must be benchmarked and validated prior to being rigorously applied for metatranscriptomics data analyses.

Although the data distributions of metagenome and metatranscriptome are regarded as of the same class, either count, normalized count, or proportion with a zero mass, the metatranscriptomics data are more sparse, or contain more zeros than metagenomics data. This higher sparsity is a likely result of some taxa, genes, or gene families not being expressed actively, but could also be attributable to some degree to less sensitive measurements of the community's transcriptome due to technical reasons. It follows that, a systematic evaluation of the performance of existing methods is warranted, so that these methods can be recommended and put to use for the analysis of metatranscriptomics data.

In this article, we focus on DE analysis of the metatranscriptome in the presence of nuisance information. The DE analysis is one of the fundamental analyses in other transcriptomics data—host bulk RNAseq and single-cell RNAseq, or scRNAseq. It often involves identifying genes whose expression levels are significantly associated with an outcome of interest (e.g., disease vs. health) after controlling for nuisance factors such as batch or block information and the host characteristics (e.g., demographics). While the metatranscriptomics DE analysis is meaningful by itself, it can provide a foundation for joint analysis of the metagenomics and metatranscriptomics data. In this article, the existence of metagenomics data is not required. Instead, the DE analysis methods for RNAseq data will be studied, accounting for the high percentage of zeros, overdispersion, possibly compositional data structures, and the presence of nuisance information.

The evaluation of methods can be done at a species, gene, or pathway level. In this paper, we focus on evaluating methods at the microbial gene level. While genes may have distinct roles within species, it is expected that the same genes across different species will have a similar functions. Furthermore, many genes are either unique to a species, or are mapped to yet-unclassified species. Metatranscriptomics provides the opportunity to measure the activity of genes, instead of inferring gene expression from microbial genomes. There have been attempts to evaluate differential abundance/composition analysis methods at the species- or the taxon-level that have included many of the commonly used pre-processing and differential abundance analyses [11, 15, 22]. However, those studies either did not consider the recently developed statistical models that were specifically designed for zero-inflated over-dispersed counts or compositional data such as MAST [23] and logistics-beta test [24] or did not evaluate their performances at the gene levels in metatranscriptomics. The number of species (typically several hundreds) and the number of genes (typically thousands, up to millions) are unique features of metatranscriptomics and metagenomics. The drawbacks become more salient when formally evaluating metatranscriptomics data analysis methods. Our paper aims at overcoming such limitations of previous attempts.

1.3 Outline

In Section 2 we list statistical tests that could be used for microbial DE analyses. In the literature of metagenomics or 16S ribosomal RNA (rRNA) data analysis, popular DA/DE analysis approaches include rank-based methods for simple experimental designs, DESeq for bulk RNAseq data[25], multi-dimensional ANCOM [26], and linear regression after log-transformation. We consider the following methods in this simulation study: 1) log-normal (LN) test, 2) logistic-beta (LB) test,

3) Model-based Analysis of Single-cell Transcriptomics (MAST), 4) DESeq2, 5) metagenomeSeq (MGS), 6) ANCOM-BC, 7) Kruskal-Wallis (KW) test, and 8) two-part Kruskal-Wallis (KW-II) test. Some of these methods are based on parametric models (LN, LB, MAST, DESeq2, MGS), some explicitly handle zero-inflation (LB, MAST, MGS, KW2), and one can handle compositional data (LB). Our simulations are more relevant than those of Weiss et al. [11], wherein most of the methods being compared do not reflect the sparse nature of metatranscriptomics data.

To understand the distributional characteristics of metatranscriptomics data, in Section 3 we introduced data collected from a genetic-epidemiologic study of early childhood oral health (the "ZOE 2.0" study) [27, 28] wherein oral microbial metatranscriptomics data were generated for approximately 300 children aged 3–5 years. We also introduced a gut microbial transcriptome dataset to study inflammatory bowel diseases [7]. Motivated by the data distributions in ZOE 2.0, we design simulations in Section 4. Our novel simulation procedure starts from defining data generative models including zero-inflated log-normal (ZILN), zero-inflated gamma (ZIG), zero-inflated negative binomial (ZINB) distributions. These distributions have the capacity of simultaneously characterizing the zero-inflation and the overdispersion of microbial gene expression. These distributional characteristics complement the recent simulation study [11]. From the distribution of the ZOE 2.0 data, a large number of baseline parameters for each generative model are identified. The disease and the batch effects are then sequentially incorporated in addition to the baseline distributions. The batch effects were inspired by the observation that different sequencing dates had different mean expression levels in the ZOE 2.0 data. However, the batch variable could be more generally defined as any set of categorical nuisance variables.

In Section 5, we present the results with respect to two main aims: 1) Goodness of fit of the different data generative models and preprocessing methods and 2) the power analysis of the DE methods via simulations. These aims are analyzed in terms of the relative gene expressions, the main focus of our study. In other words, each gene's expression levels in RPKs are summed (or marginalized) over all species before being analyzed. However, other aspects of microbial activities—the gene expression of a combination of a species and a gene ("gene-species joint data") and the gene expression of a species aggregated over all genes ("species marginal data")—are also considered.

We first show the goodness of fit results of the generative models to the data sets after a certain pre-processing. The results suggest that the log-normal distribution fits reasonably well to the ZOE 2.0 data with the TPM transformation. Second, in simulations, the LB tests showed the highest power while controlling for the type-I error in large sample sizes. Also MGS has comparably high power with a well-controlled type-I error. MAST and ANCOM-BC often suffers from inflated type-I error even with a large sample size.

In Section 6, we apply the statistical analysis methods that show the highest performance to the sizeable metatrascriptomics dataset generated in the ZOE 2.0 pediatric oral health study. We identify several genes of which the expression is associated with the early childhood caries (ECC) phenotype [29]. The paper concludes with a discussion of considerations for future studies in Section 7.

2 Differential expression analysis methods

2.1 Overview

The eight DE analysis methods are evaluated in this simulation study are Log-normal test (LN), Logistic Beta test (LB) [24], Model-based Analysis of Single-cell Transcriptomics (MAST) [23], DESeq2 [25], metagenomeSeq (MGS) [22], ANCOM-BC [30], Kruskal-Wallis test (KW), and two-part Kruskal-Wallis test (KW-II). Most methods are tailored so that they can control for batch effects while testing associations between gene expression and phenotypes of interest.

Linear discriminant analysis Effect Size (LEfSe) [31] was developed to identify species or genes that are likely to differentiate two or more groups in terms of the relative abundance of these features. Because it does not result in a statistical testing *per se*—the p-value is not available—we do not include it in this simulation study. The differential ranking (DR) method [32] and ANCOM [26] were not analyzed for the same reason. Although MGS and ANCOM-BC does not control the batch effects as well, we included them in the simulations due to its high relevance in the metatranscriptomics research.

The scope of our simulations is restricted to tests of differential expression of individual taxa, gene, or gene family level. On the other hand, there is another group of methods that test the global effect of the whole microbiome, which will not be covered in this simulation study: PERMANOVA [33], MiRKAT [34], aMiSPU [35], and LDM [36].

Before we present each of eight methods in detail in the following subsections, we introduce notational conventions and describe the screening procedures. Throughout this paper, $Y_{i,g}$ denotes the expression level for the gth gene in the ith cell, $\mathbf{X}_i \equiv (1, X_i^D, X_i^B)$ denotes the ith row of the design matrix, or a vector containing the intercept term, a binary disease status, and a binary batch indicator. Different models abusively use the same notation for parameters, as long as there is no confusion; e.g. regression coefficients β_g are commonly used either in the LN model or the LB model, but they are shorthand for β_g^{LN} and β_g^{LB}, respectively.

Before each test, genes are screened out if they are expressed in only a few participants (the smaller of 2% of the samples and 10 samples). The rationale for such screening is twofold. First, models are not mathematically estimable, when the sample size is less than the number of parameters in the model. Second, even if they are estimable, many large-sample-based inferential procedures have non-negligible finite-sample bias when they are implemented in a small sample or when a gene expression is rare [37]. These issues become apparent when 1) the tests that involve logistic regression encounter rare events or 2) the two-part models where non-zero statistics are faced with only a small number of observations with nonzero expression. As evidenced in later sections, all methods except KW fall into either of these two scenarios. To allow for fair comparisons, we apply the same screening rule for each test.

2.2 Log-normal test

The Log-normal (LN) test relies on the assumption that the log-transformed expression is normally distributed as in (1). A small positive constant (c) is added to the gene expression to ensure that the log-transformed values are within a feasible range. In this simulation study 1 is uniformly added to expression levels ($c = 1$).

$$\log_2(Y_{i,g} + c) \sim \mathcal{N}(\mu_{i,g}, \sigma_g), \tag{1}$$

where $\mu_{i,g} \equiv \boldsymbol{X}_i^\top \beta_g$ with $\beta_g \equiv (\beta_g^0, \beta_g^D, \beta_g^B)^\top$.

The null and the alternative hypotheses for the gth gene are

- H_0: $\beta_g^D = 0$ and
- H_1: $\beta_g^D \neq 0$.

The test statistic for the gth gene is $T_g^{LN} = \left(\frac{\hat{\beta}_g^D}{se(\hat{\beta}_g^D)}\right)^2$ and follows a χ_1^2 distribution under the null hypothesis asymptotically. The test rejects the null hypothesis if the test statistic is larger than $\chi_1^2(1-\alpha)$, or the $(1-\alpha)$th quantile of the χ^2 distribution with one degree of freedom, where α is the significance level. Alternatively, the individual p-values are obtained as $p_g = 1 - F_{\chi_1^2}(T_g^{LN})$, where $F_d(t)$ is the distribution function of d evaluated at t. The genes with p-values less than α are declared to have a statistically significant association with disease. This test is simply an analysis of covariance (ANCOVA) with an appropriately log-transformed dependent variable, and is easily implemented in most statistical software packages. The testing procedure, after obtaining a test statistic and the corresponding null distribution (e.g., p-values and rejection regions), is identical for rest of the methods, hence will be omitted unless needed.

2.3 Logistic Beta test

The Logistic Beta model (LB) models relative expressions, $R_{i,g} \equiv Y_{i,g}/\sum_{h=1}^{G} Y_{i,h}$, instead of absolute expressions, $Y_{i,g}$. Because of the sum-to-one constraint of relative expressions, tests based on relative expression are structurally dependent. However, in microbiome data analyses, the number of tested genes is usually large enough and thus the dependence induced by the compositional structure is negligible.

The LB model is formulated [24] as:

$$R_{i,g} \sim \mathcal{LB}(\pi_{i,g}, \mu_{i,g}, \phi_g), \tag{2}$$

where $\pi_{i,g} = \text{expit}(\boldsymbol{X}_i^\top \gamma_g)$ with $\gamma_g = (\gamma_g^0, \gamma_g^D, \gamma_g^B)^\top$, $\mu_{i,g} = \text{expit}(\boldsymbol{X}_i^\top \beta_g)$ with $\beta_g \equiv (\beta_g^0, \beta_g^D, \beta_g^B)^\top$, ϕ_g denotes the dispersion parameter such that $var(R_{i,g}|R_{i,g} > 0) = \mu_{i,g}(1-\mu_{i,g})\phi_g$, and $\text{expit}(\cdot) := \frac{\exp(\cdot)}{\exp(\cdot)+1}$.

Note that this model can be decomposed into two orthogonal models:

$$1(R_{i,g} = 0) \sim \text{Bernoulli}(\pi_{i,g}), \qquad R_{i,g}|R_{i,g} > 0 \sim \text{Beta}(\mu_{i,g}, \theta_g), \tag{3}$$

where $1(\cdot)$ is the indicator function, $\mu_{i,g}$ is the mean of the Beta random variable and θ_g is the dispersion parameter. Orthogonality means that the estimate of $\pi_{i,g}$ and those of $\mu_{i,g}$ and θ_g are independent. Consequently, the test statistic can be obtained from these two separately estimated models. The maximum likelihood estimators (MLE) are used for estimation and an `R` package `gamlss` [38] was used for simulation in this study.

The null and the alternative hypotheses for the gth gene are

- H_0: $\beta_g^D = \gamma_g^D = 0$ and
- H_1: Either $\beta_g^D \neq 0$ or $\gamma_g^D \neq 0$.

Either a Wald-type or a likelihood test statistic can be used to test these hypotheses. Because they are asymptotically equivalent, here we only present a Wald test statistic:

$$T_g^{LB} = \left(\frac{\hat{\beta}_g^D}{se(\hat{\beta}_g^D)} \right)^2 + \left(\frac{\hat{\gamma}_g^D}{se(\hat{\gamma}_g^D)} \right)^2, \tag{4}$$

which follows a χ_2^2 distribution under the null hypothesis asymptotically.

If only one of the two parts of LB is estimable, the test statistic is constructed based on the estimable component only and the reference distribution is χ_1^2; e.g. when only the logistic model is estimable, $T_g^{LB} = \left(\frac{\hat{\gamma}_g^D}{se(\hat{\gamma}_g^D)} \right)^2$. The same approach was followed for the other two-part tests, including MAST and the two-part Kruskal-Wallis test.

2.4 MAST

The“Model-based Analysis of Single-cell Transcriptomics” (MAST) [39] was proposed specifically for differential expression analysis of scRNAseq data. This model, composed of a logistic regression model and a conditional log-normal model, regularizes parameter estimation and utilizes estimated cellular detection rates (CDR) as covariates as defined below. The model was designed to deal with zero-inflation which is driven by both technical and biological variabilities in scRNAseq data. Though zeros in microbiome sequencing data are believed to be generated mostly by biological reasons, the proportion of zeros is usually greater than that of conventional single-part parametric models such as Poisson, negative binomial, and log-normal. Thus, it is feasible to interrogate the performance of MAST in the context of microbiomal transcriptomics analysis.

The models in MAST can be summarized as

$$1(Y_{i,g} = 0) \sim \text{Bernoulli}(\pi_{i,g}), \quad \log_2(Y_{i,g} + 1) \mid Y_{i,g} > 0 \sim \mathcal{N}(\mu_{i,g}, \sigma_g^2), \tag{5}$$

where $\pi_{i,g} = \text{expit}(\boldsymbol{X}_i^\top \gamma_g)$ with $\gamma_g \equiv (\gamma_g^0, \gamma_g^D, \gamma_g^B, \gamma_g^C)^\top$, $\mu_{i,g} = (\boldsymbol{X}_i^\top \beta_g)$ with $\beta_g \equiv (\beta_g^0, \beta_g^D, \beta_g^B, \beta_g^C)^\top$, $\mathbf{X}_i \equiv (1, X_i^D, X_i^B, X_i^C)$, and $X_i^C = \frac{1}{n} \sum_{i=1}^n 1(Y_{i,g} > k)$ is the CDR of the ith subject for background expression level k. In this simulation we set $k = 0$.

The parameters are estimated using a Bayesian framework, where γ_g is regularized under weak informative prior and $1/\sigma_g$ is regularized using empirical Gamma prior. An `R` package `mast` is available [23].

The null and the alternative hypotheses for the gth gene are

- H_0: $\beta_g^D = \gamma_g^D = 0$ and
- H_1: Either $\beta_g^D \neq 0$ or $\gamma_g^D \neq 0$.

Either a Wald-type or a likelihood test statistic can be used to test these hypotheses. The Wald statistic is

$$T_g^{MAST} = \left(\frac{\hat{\beta}_g^D}{se(\hat{\beta}_g^D)} \right)^2 + \left(\frac{\hat{\gamma}_g^D}{se(\hat{\gamma}_g^D)} \right)^2, \tag{6}$$

with χ^2_2 as its asymptotic null distribution. The testing procedure is exactly the same as that of the LB test once the coefficients and their standard errors are estimated.

2.5 DESeq2

The DESeq2 [25] method is currently widely used for differential expression of RNAseq data. The underlying model of DESeq2 is a negative binomial distribution and it uses empirical Bayes for regularization.

The DESeq2 model can be summarized as

$$Y_{i,g} \sim \mathcal{NB}(\mu_{i,g}, \theta_g), \tag{7}$$

where $\mu_{i,g} = s_{i,g}\nu_{i,g}$ is the mean parameter, θ_g is the dispersion parameter, $s_{i,g}$ is the size factor, $\nu_{i,g} = \exp(\boldsymbol{X}_i^\top \beta_g)$ with $\beta_g \equiv (\beta_g^0, \beta_g^D, \beta_g^B, \beta_g^C)^\top$, and $\mathbf{X}_i \equiv (1, X_i^D, X_i^B, X_i^C)$. The size factor is the parameter with which we adjust the sequencing depth. In this simulations we use the median-of-ratios method [40].

The parameters are estimated using maximum likelihood estimation and then θ_g and β_g^D are regularized using an empirical Bayes approach. An `R` package `DESeq2` is available.

The null and the alternative hypotheses for the gth gene are

- H_0: $\beta_g^D = \gamma_g^D = 0$ and
- H_1: Either $\beta_g^D \neq 0$ or $\gamma_g^D \neq 0$.

A Wald test is used to test these hypotheses. The Wald statistic is given as

$$T_g^{DESeq2} = \left(\frac{\hat{\beta}_g^D}{se(\hat{\beta}_g^D)}\right)^2, \tag{8}$$

with χ^2_1 as its asymptotic null distribution. The testing procedure is exactly the same as that of LB test, once the coefficients and their standard errors are estimated.

Because DESeq2 cannot accomodate high zero proportions, an extension was recently developed to enable the modeling of a greater number of zeros in the scRNAseq context [41]. In this modified DESeq2 method, namely DESeq2-ZINBWaVE, first the zero-inflation parameter is estimated using the model,

$$Y_{i,g} \sim \mathcal{ZINB}(\mu_{i,g}, \theta_g, \pi_{i,g}), \tag{9}$$

and each observation is assigned a weight of the posterior probability of non-zero-inflation, $\frac{(1-\pi_{i,g}) f_{ZINB}(y_{i,g};\mu_{i,g},\theta_g,0)}{f_{ZINB}(y_{i,g};\mu_{i,g},\theta_g,\pi_{i,g})}$, where f_{ZINB} is the corresponding density of the ZINB distribution. For the size factor estimation in DESeq2-ZINBWaVE, we use the positive counts method. Then the conventional DESeq2 method is applied, as described earlier, including the weights. Whenever there is no ambiguity, “DESeq2” refers to the original method and “DESeq2-ZINBWaVE” to its extension.

2.6 metagenomeSeq

MetagenomeSeq (MGS) is a differential abundance analysis method for metagenomics data [22] and the corresponding bioconductor package, `metagenomeSeq` is available. MGS assumes zero-inflated log normal distribution. Furthermore, MGS uses an empirical Bayes shrinkage method for parameter estimation. Hence, MGS shares common modeling approaches with MAST; however, the two approaches are different in a few aspects. First, MAST uses CDR as a controlling factor in the model while MGS does not. Second, MAST provides tests on two parts of the model; i.e., two p-values are obtained from the zero-inflation part and the log-normal part in MAST. However, in MGS, after estimating the two-part model parameters, only the log-difference of the marginal mean is tested and a single p-value is given [42]. In our simulations, the `fitFeatureModel` function in the `R` package `metagenomeSeq` is used for implementation [43]. Although the MGS test can account for batch effects mathematically, the current `metagenomeSeq` software does not allow batch variables in the model. Thus, only results without batch effects will be reported in the simulation study in Section 4.

2.7 ANCOM-BC

ANCOM-BC is a differential abundance analysis method for metagenomics data [30]. It shares the philosophy of its predecessor, ANCOM [26], in that it models the ratio of abundances between taxa. However, unlike ANCOM which is a rank-based approach, ANCOM-BC specifies the test statistic and its associated p-value for a large sample. In ANCOM-BC, the observed abundance $Y_{i,g}$ is assumed to be a realization of the unknown abundance $U_{i,g}$ of the whole ecosystem from where the sample is taken with possibly different sampling fraction η_i for each sample. In other words, $E[Y_{i,g}|U_{i,g}] = \eta_i U_{i,g}$, where $U_{i,g}$ is a random variable with mean θ_g^D or θ_g^H, depending on the membership of the sample i to the disease (D) or health (H) group. Of note, ANCOM-BC is not limited to two-group problems but are designed for multi-group problems. Then it formulates $\log Y_{i,g} = \log \tilde{\eta}_i + \log \theta_{i,g} + \epsilon_{i,g}$, where $\tilde{\eta}_i$ is a slightly-redefined sampling fraction parameter due to the log-transformation, and $E[\epsilon_{i,g}] = 0$.

The hypotheses of ANCOM-BC are

- H_0: $\log \theta_g^D = \log \theta_g^H$ and
- H_1: $\log \theta_g^D \neq \log \theta_g^H$,

which are tested by the test statistic,

$$T_g^{ANCOM-BC} = \frac{\widehat{\log \theta_g^D} - \widehat{\log \theta_g^H} - \widehat{\log \tilde{\eta}}}{\sqrt{\{\hat{\sigma}_g^D\}^2 + \{\hat{\sigma}_g^H\}^2}},$$

where $\widehat{\log \theta_g^A}$ is the estimates of $\log \theta_g^A$, $\{\hat{\sigma}_g^A\}^2$ is the mean squared error for each group $A = D, H$, and $\widehat{\log \tilde{\eta}}$ is the estimate of the bias, $\log \tilde{\eta} \equiv E[\log \theta_g^D - \log \theta_g^H]$. The statistic follows the standard normal distribution per the large sample theory, and the authors defined a small sample version of the statistic of which distribution was not defined.

ANCOM-BC does a further procedure of detecting "the structural zero" which is defined to be the absence of a certain taxon in a specific group that is present in another group. Once the structural zero is detected, ANCOM-BC declares that the taxon is differentially abundant, giving $T_g^{ANCOM-BC} = \infty$ and p-value $= 0$. However, since such procedure often inflates the type-I error significantly in this simulation study, we add another version of ANCOM-BC that declares those sturctural zeros inconclusive (i.e., p-value $=$ NA). We report the simulations results for second version as “ANCOM-BC2” and disclose the results of the original version, denoted as “ANCOM-BC1,” in the Supplementary Materials.

2.8 Kruskal-Wallis test

The Kruskal-Wallis (KW) test is equivalent to one-way analysis of variance (ANOVA) on ranks. KW is equivalent to Wilcoxon’s rank sum (WRS) test, or Wilcoxon-Mann-Whitney test, for two sample problems and can accomodate comparisons of more than two samples [44]. Although the prototypical KW test was designed without consideration of covariates, it can be modified to account for possible batch effects [45]:

$$T_g^{KW} \equiv (n-1)\frac{\sum_{i=1}^{n}(\bar{r}_g^{db} - \bar{r}_g^{\cdot b})^2}{\sum_{i=1}^{n}(r_{i,g} - \bar{r}_g^{db})^2}, \tag{10}$$

where $r_{i,g} := \sum_{j=1}^{n}\{1(Y_{i,g} > Y_{j,g}) + \frac{1}{2}(Y_{i,g} = Y_{j,g})\} + 1$ is the rank of the ith subject’s gth gene among n subjects, $\bar{r}_g^{db} := \frac{\sum_{i=1}^{n} r_{i,g} 1(X_i^D = d, X_i^B = b)}{\sum_{i=1}^{n} 1(X_i^D = d, X_i^B = b)}$, $\bar{r}_g^{d\cdot} := \frac{\sum_{i=1}^{n} r_{i,g} 1(X_i^D = d)}{\sum_{i=1}^{n} 1(X_i^D = d)}$, and $\bar{r}_g^{\cdot\cdot} := \frac{1}{n}\sum_{i=1}^{n} r_{i,g}$.

The null and the alternative hypotheses for the gth gene are

- H_0: The ranked expression levels are independent of the phenotypic outcome controlling for batch effects,
- H_1: The complement of H_0.

The exact and approximate distributions of the statistic under the null hypothesis can be obtained through analysis or resampling [45]. However, when disease and batch strata are large, the statistic converges to χ^2_1. Based on the null distribution, p-values are obtained for each gene.

For implementation of this test, an `R` function `coin::kruskal_wallis()` [46] is available. The `coin` package function allows only a single batch variable.

2.9 two-part Kruskal-Wallis test

Nonparametric tests such as KW and Wilcoxon’s rank-sum (WRS) test have minimal distributional assumptions. The lack of model-induced information often results in lack of power. While zero-inflation is a well-known characteristic of microbiome sequencing data, explicitly modeling the proportion of zeros can enhance the power in detecting differential expression. This additional assumption can be integrated into the nonparametric models using two-part model framework [47]. Of note, the LB test and the MAST are also two-part models but they are fully parametric. Nonparametric two-part models have been used in other ’omics applications [48] and microbiome [49] data analysis. The binary part of these nonparametric models has been modeled using a conventional proportion test where no covariates are allowed. To incorporate covariate information in the binary model, a logistic regression model can be used. The KW or the WRS test can be used as the nonzero model–to allow for the inclusion of covariates in the model, the modified KW test can be used. In this paper we combine a logistic regression model and a KW test and name it two-part KW test. The binary component of the model is the same as that of LB model, i.e. Equation (3). The nonzero component's test statistic is derived based only on subjects with non-zero gene expressions and has the same formula given in KW test, i.e. Equation (10).

The p-values can be obtained by combining two χ_1^2 statistics derived from each component:

$$T_g^{KWII} := W_g^a + W_g^b,$$

where W_g^a is either a Wald test statistic ($\left(\frac{\hat{\beta}_g^D}{se(\hat{\beta}_g^D)}\right)^2$) or a likelihood ratio statistic (two times the difference of log-likelihood of logistic regression models), and W_g^b has the same form as Equation 10. The test statistic, T_g^{KWII} follows a χ_1^2 distribution under the null hypothesis asymptotically.

3 Description of the three metatranscriptomics datasets

To understand and characterize the distributional features of metatranscriptomics data, we leverage three datasets made available by two recent studies involving the human microbiome. The first two datasets (namely, ZOE 2.0 and ZOE-pilot) were generated in a molecular epidemiologic study of early childhood caries (ECC; defined as dental cavities in children under the age of 6) [29] called Zero Out Early Childhood Tooth Decay (ZOE) [28, 27]. In that study, the association between the supragingival oral microbiome and the prevalence of clinically-determined ECC is investigated. The third dataset (namely, the IBD data) was generated in the context of a recent study of the gut microbial ecosystem and its association with Inflammatory Bowel Diseases (IBD) [7]. We base our analyses mainly on the largest ($n = 300$) oral microbiome dataset, or the ZOE 2.0 data, and use the other two datasets for the purposes of validation.

3.1 The pediatric dental caries datasets

One of the main aims of the ZOE project is to understand the biological basis of ECC, including the human genome and the oral microbiome. To-date, approximately 5% of the parent cohort ("ZOE 2.0", 300/6,404) has been carried forward metagenomics, metatranscriptomics, and metabolomics analyses [27]. In addition, 118 participants from the same population were included in a pilot study ("ZOE-pilot") that included identical phenotyping and biofilm sequencing procedures. In sum, dental biofilm metatranscriptomics analyses have been done to-date on 418 children ages 3–5 [27]. As noted above, the ZOE data were harvested in two different periods. First, the current ZOE 2.0 samples of 300 children were sequenced in 2018-19 and the 118 samples of the ZOE-pilot cohort were sequenced in 2018-19. Further, within ZOE 2.0, 53 samples were sequenced in May 2018 and 249 were sequenced in November of 2019, respectively. Importantly, the average sequencing depth varied significantly across the sequencing dates, and thus the dates are considered as batches in the parameter selection procedure in Section 4.3. Similarly, batch effects were evident in the ZOE-pilot data: 60 samples were sequenced in June and an additional 58 samples were sequenced in July of 2017, respectively. Three and two subjects were excluded for further analyses due to no or significantly low expression levels in the ZOE 2.0 and the ZOE-pilot data, respectively. For the purposes of this analysis, ECC was defined as a dichotomous trait, healthy or diseased, based on modified International Caries Detection and Assessment (ICDAS) criteria [28]. ECC prevalence was similar in the two ZOE waves, i.e., ZOE 2.0: 49% (147/297) and ZOE-pilot: 50% (58/116). A detailed microbiome analysis protocol for this study has been reported recently [27].

Adapter-trimmed, quality-controlled, demultiplexed Illumina HiSeq sequencing reads were aligned against the human hg19 reference to eliminate host derived reads. Generally speaking, the alignment and data pre-processing followed the procedure described previously [27]. For details, estimates of taxonomic composition, gene family, path abundance, and path coverage were produced from the remaining reads using HUMAnN2 [50]. The resulting reads were scaled into reads-per-kilobase (RPK). We considered additional pre-processing methods including transcript-per-kilobase-million (TPM) and arcsine in Section 5.

The total number of gene-species combinations in the ZOE 2.0 metatranscriptome is 535,299; there are 204 distinct species, and 402,937 distinct genes. In the ZOE-pilot sample, there are 439,872 gene-species, 185 distinct species, and 342,004 distinct genes. Total RPKs per sample is on average 13,053,428 in ZOE 2.0 and 2,815,749 in ZOE-pilot. The RPKs are rescaled by dividing by the total RPKs per sample and then multiplying by 4.0 million in ZOE 2.0 and 3.4 million in ZOE-pilot, to make the total expression level for each subject to be 10 times the number of genes. This is a scaled version of TPM-normalized data. In this article, for notational convenience, this scaled version of TPM is referred to as TPM.

There were high proportions of zero gene expressions in both the ZOE 2.0 (80.4%) and the ZOE-pilot (87.9%) metatranscriptomics data. These high zero proportions are comparable and actually higher than what is encountered in the corresponding metagenomics data (75% in ZOE 2.0 and 68% in ZOE-pilot), as illustrated in Figure 1. While a significant number of genes in the metagenomics data are not sparse—one ninth in the ZOE 2.0 (one sixth in ZOE-pilot) of all genes have zero proportion smaller than 20%, virtually all genes, or 94% (97%) are sparse in the ZOE 2.0 (ZOE-pilot) metatranscriptomics data. Specifically, 54% (43%) of genes have $\geq$ 90% zero proportions in metagenomics compared to 59% (71%) in the metatranscriptomics data of the ZOE 2.0 (ZOE-pilot) data.

3.2 The inflammatory bowel diseases dataset

The second dataset that we use in this study is one generated in the context of IBD multi-omics research wherein the association between IBD and the gut microbial ecosystem was studied [7]. The investigators followed 132 subjects for a one-year period obtaining repeated measurements of multi-omics components including fecal metatranscriptomes over multiple time points. For metatranscriptomes, a modified RNAtag-seq protocol was used to create Illumina cDNA libraries which were sequenced on the Illumina HiSeq2500 platform yielding approximately 13 million paired end reads. In this IBD study, the authors have generated the taxonomic and functional profiles (http://huttenhower.sph.harvard.edu/biobakery). Reads were mapped the human genome. Taxonomic profiles of shotgun metagenomes were generated using MetaPhlAn. Functional profiling was performed by HUMAnN2 to quantify gene presence and abundance on a per-species basis (UniRef90s), for both metagenomics and metatranscriptomics. To ensure a reasonable read depth in each sample, the authors only used samples (metagenomes and metatranscriptomes) with at least 1 million reads (after human filtering) and at least one non-zero microbial abundance detected by MetaPhlAn. The dataset includes a total of 1,595 metagenomic and 818 metatranscriptomic samples.

We focus on the cross-sectional features of the metatranscriptomics data distribution, and thus we only examined the baseline information, or the first visit data, including a sample of 104 participants. We further dichotomized participants' disease status as IBD (i.e., 50 Crohn's disease and 26 ulcerative colitis cases) versus non-IBD (i.e., 28 'control' participants). Clinic location was considered as a batch effect in that study, and thus was employed in our analyses after dichotomization (the pediatric versus the adult cohorts). The total number of gene-species combinations in the IBD dataset was 42,688, including 235 distinct species, and 1,629 distinct genes, among samples that had at least 1 million reads and at least one non-zero microbial abundance captured by MetaPhlAn2. The data are publicly available in a compositional format, where the gene expression sums up to one over genes and species for each measurement of a subject.

The average proportion of zeros per gene in the IBD metatranscriptomics data is 96.3%, while that in the metagenomics data is 87.8% consistent with the trend of higher zero proportions in metatranscriptomics data over metagenomics data. Figure 1 illustrates the higher zero proportion in the metatranscriptomes over the metagenomes: 91% of genes have zero proportion $\geq 90\%$ in the metatranscriptomics data, compared to 69% in the metagenomics data.

3.3 Data scaling and transformation

It is widely acknowledged that the DE analysis in metatranscriptomics depends not only on the DE methods but also on pre-processing of the data. For count data, scaling and transformation are commonly used as part of data normalization to make the data more comparable across samples and/or taxa or to remove distributional irregularities such as skewness. Reads-per-kilobase (RPK), transcripts-per-kilobase-million (TPM), rarefying, and upper-quartile log-fold change normalization [51] are frequently used as the scaling techniques. The examples of transformation include arcsine, logarithm, and variance stabilizing transformation (VST) [40] among many others.

The strengths and the shortcomings of each scaling and transformation approach have been previously presented and discussed in the literature [52, 13], and also some simulation studies have been done to compare the methods in the differential abundance testing context [11]. However, in this paper, we do not pursue a comprehensive comparison of the scaling and transformation methods. Rather, we only consider the three widely used methods, RPK, TPM, and arcsine, because the DE methods evaluation is the main aim of our paper and, also, those scaling and transformation methods provide reasonably good distributional results in the example data.

Among the three pre-processing methods, TPM is defined as RPK divided by the sample sum of RPKs times a constant ($TPM_{i,g} = c\frac{RPK_{i,g}}{\sum_{g=1}^{n_{genes}} RPK_{i,g}}$) where the constant, c, was chosen to reflect the actual scale of the RPKs, or $c = 5(20)$ million in the ZOE-pilot (2.0) study. Arcsine-transformation is defined as $c\frac{2}{\pi}\arcsin(\sqrt{TPM/c})$. Note that the Beta distribution is for compositional data and thus RPK and TPM data are equivalent for the Beta distribution. The IBD data are available only in a compositional form, and thus, we do not consider the RPK-form of data.

4 Simulation design

4.1 Overview

To comprehensively evaluate the performance of the available tests, we considered multiple scenarios defined by the following generative models with a nested factorial design. Three data generative model classes were used: zero-inflated log-normal (ZILN), zero-inflated negative binomial (ZINB) and zero-inflated gamma (ZIG) models. Each of the generative models is characterized by three factors: 1) baseline distribution, 2) disease effects, and 3) batch effects (without interaction with disease), and they are defined in Sections 4.2 and 4.3. The factorial elements of the simulations are selected based on the ZOE 2.0 data. For each generative model, we estimate the parameters of the transformed gene expression of randomly chosen 300 genes in that dataset. A set of parameter values for each factor is chosen so that it reasonably covers the distribution of the estimates.

We first obtain the baseline distribution of parameters by estimating them marginally in dental health and disease (H: non-ECC and D: ECC) and sequencing groups (i.e., batches). Next, we obtain the distribution of the estimated disease and batch effect parameters. Sets of parameters were chosen for this simulation study so that they reasonably represent the real data distributions of parameter estimates. Details regarding parameter selections are described in the corresponding subsections of Section 4.3. Once the data distribution is defined, we generate random samples of a small ($n = 80$) and a large ($n = 400$) size, where all four subgroups of disease-batch combinations are equally sized and there are $n_{\text{gene}} = 10,000$ genes. Then we apply all tests listed in Section 2, and calculate the average rejection rate of each test at the 5% significance level. However, to avoid possibly high variability of averaged rejection rate of MAST, DESeq2, MGS, and ANCOM-BC where individual tests for each of n_{gene} genes are done non-independently, we ran 10 sets of each of those three tests using additional data replicates and this way obtain the average rejection rates. Simulations are done in `R` 4.0.3 and the code is available at https://github.com/Hunyong/microbiome2020.

4.2 Generative models

Three generative models are considered: zero-inflated log-normal (ZILN), zero-inflated gamma (ZIG), and zero-inflated negative binomial (ZINB). We do not include the zero-inflated beta (ZIB) distribution, as only a few methods, such as the LB test, model relative gene expressions or abundances rather than their absolute quantities. Furthermore, because ZIB can be considered a compositional transformation of independent ZIG distribution, ZIG-based results should serve as a good proxy for ZIB-based simulations.

4.2.1 Generative model 1 - ZILN-based

The zero-inflated log-normal model is a mixture of log-normal distribution with a point mass at zero. The density is given as

$$f_{ZILN}(y) = \pi \; 1(y=0) + (1-\pi) \; \varphi(y, \mu, \mu\theta) 1(y>0), \tag{11}$$

where $\varphi(x, \mu, \sigma^2)$ is the log-normal density at x with mean μ and variance σ^2, π is the zero-inflation parameter, or $\pi \equiv \Pr(Y=0)$, μ is the non-zero mean parameter (i.e. $\mu \equiv E[Y|Y>0]$), and θ is the over-dispersion parameter so that $var[Y|Y>0] = \mu^2\theta$

4.2.2 Generative model 2 - ZINB-based

ZINB is an extension of the negative binomial distribution and is widely used to model count data with excess zeros. In many real world applications, if more zeros are observed than the negative binomial distribution assumes, the zero-inflated negative binomial distribution is suitable and has in fact become one of the most commonly used methods in count data analysis [53]. This is true for omics data analysis including scRNAseq [54, 41] and microbiome data [55, 56].

ZINB without covariates has three parameters, $(\mu, \theta, \pi)^\top$, with the following density:

$$f_{ZINB}(y) = \pi \; 1(y=0) + (1-\pi) \; \binom{y + \frac{1}{\theta} - 1}{y} \frac{(\mu\theta)^y}{(1+\mu\theta)^{y+1/\theta}}, \tag{12}$$

$y = 0, 1, 2, ...$, where π is the zero-inflation parameter, μ is the mean parameter assuming no zero-inflation (i.e. $E[Y] = \mu(1-\pi)$), and θ is the over-dispersion parameter such that $var[Y] = \mu^2\pi(1-\pi) + (1-\pi)(\mu + \mu^2\theta)$.

We use the same notation $\xi \equiv (\mu, \theta, \pi)$ for each generative model as long as there is no ambiguity, and use a superscript denoting the model if distinction is needed.

4.2.3 Generative model 3 - ZIG-based

The zero-inflated Gamma model is a mixture of a Gamma distribution and a point mass at zero. The density is given as

$$f_{ZIG}(y) = \pi \; 1(y=0) + (1-\pi) \; \frac{y^{\mu/\theta - 1} e^{-y/\theta}}{\Gamma(\mu/\theta)\theta^{\mu/\theta}} 1(y>0), \tag{13}$$

where π is the zero-inflation parameter or $\pi \equiv \Pr(Y=0)$, μ is the non-zero mean parameter (i.e. $\mu \equiv E[Y|Y>0]$), and θ is the over-dispersion parameter so that $var[Y|Y>0] = \mu^2\theta$.

4.3 Model parameters

4.3.1 Baseline parameters

A set of baseline parameters uniquely defines a null distribution where there is neither disease nor batch effects. Based on each of these baseline distributions, the disease and/or batch effects are added to form alternative distributions.

The ZILN parameters were estimated (Figure 2A) for the 300 randomly chosen genes in the ZOE 2.0 dataset for each of the disease and batch subgroups. The method of moments was used for estimation. The first, the second, and the third quartiles of the μ estimates are 7.4, 19.7, and 52.6, respectively. Those for the θ estimates are 0.7, 1.2, and 1.8. Those for the π estimates are 0.34, 0.64, and 0.83. Based on the parameter distribution, we selected sets of baseline parameters for the ZILN model as in Table 1. The parameter estimates for the ZIG model are identical to those of the ZILN model and for this reason are not presented. ZINB model parameter estimates are provided in Supplementary Section 3.1 of the Supplementary Material.

To add to our understanding of metatranscriptomics data distributions generated under realistic conditions, we followed the same procedures to estimate the parameters using the ZOE-pilot and the IBD data. Because the ZILN model is the main focus of our simulation study and the expression of the IBD data was only provided in a compositional form, we use these validation data to estimate the ZILN model parameters only. In the ZILN model, μ is the only parameter that is affected by scale transformations and most test results are thus invariant to scale transformations except for the NB- and the ZINB-based tests. These estimated parameters from the ZOE-pilot and the IBD data are presented in Supplementary Section 1. In Supplementary Figure 1A and 2A, the estimated parameters of the gene expression distribution for genes in the ZOE-pilot and the IBD data have a range that overlaps reasonably with Figure 2A and the sets of parameters in Table 1.

Of note, in this paper we focus on total gene expression for each gene aggregated over all the species in the sample. We further consider parameters for i) expression of each gene-species combination (i.e., the joint data) and ii) species expression marginally across genes (i.e., the species marginal data). The corresponding distributions are provided in Supplementary Section 2 of the Supplementary Material. The distributions from the gene-species joint data and the species marginal data are reasonably covered by the parameter sets chosen in this section for all the settings except the ZINB model for the marginal species data, where the parameters are often either not estimable or outlying. In other words, the results are generalizable to the other aspects of data and in most settings, except for the species data with the ZINB model.

4.3.2 Disease effects

For each of the baseline distributions, disease effects are further considered to construct the alternative distributions. Let $\delta \equiv (\delta_\mu, \delta_\theta, \delta_\pi)$ denote the disease effects such that $\log \mu_D = \log \mu + \frac{1}{2}\delta_\mu$, $\log \theta_D = \log \theta + \frac{1}{2}\delta_\theta$, and $\log \frac{\pi_D}{1-\pi_D} = \log \frac{\pi}{1-\pi} + \frac{1}{2}\delta_\pi$, where $\xi_D \equiv (\mu_D, \theta_D, \pi_D)$ is the parameter for the diseased group. We simply denote such operation as $\xi_D = g(\xi, \delta)$. The parameter for the healthy group is $\xi_H \equiv (\mu_H, \theta_H, \pi_H) = g(\xi, -\delta)$.

The disease effect estimates for the ZILN model are estimated from randomly chosen 300 genes from the ZOE 2.0 dataset assuming that there are no batch effects. These estimates are presented in Figure 2B. The quartiles of the δ_μ estimates are 0.1, 0.2, and 0.5 in the order. Those for the δ_θ estimates are 0.3, 0.5, and 0.9. Those for the δ_π estimates with finite values are 0.2, 0.3, and 0.6. Based on the

No.	μ	θ	π
B01-03	1, 10, 50	0.5	0.3
B04-06	1, 10, 50	2.0	0.3
B07-09	1, 10, 50	0.5	0.6
B10-12	1, 10, 50	2.0	0.6
B13-15	1, 10, 50	0.5	0.65
B16-18	1, 10, 50	2.0	0.65
B19-21	1, 10, 50	0.5	0.7
B22-24	1, 10, 50	2.0	0.7
B25-27	1, 10, 50	0.5	0.75
B28-30	1, 10, 50	2.0	0.75
B31-33	1, 10, 50	0.5	0.8
B34-36	1, 10, 50	2.0	0.8
B37-39	1, 10, 50	0.5	0.85
B40-42	1, 10, 50	2.0	0.85
B43-45	1, 10, 50	0.5	0.9
B46-48	1, 10, 50	2.0	0.9
B49-51	1, 10, 50	0.5	0.95
B52-54	1, 10, 50	2.0	0.95

Table 1: Baseline ZILN parameters

parameter estimates, we select sets of disease effects for ZILN model as in Table 2. Note that the π effect of scenario D4 is -1 to maintain consistency of the direction of the effects. The parameter estimates for the ZINB are provided in Supplementary Figure 7B. The corresponding ZILN parameter estimates for the gene-species joint data and the species marginal data are presented in Supplementary Figure 3B and 5B. The estimated ZILN parameters of the gene expression distribution for genes, gene-species, and species have a range that overlaps reasonably with Figure 2B and the sets of parameters in Table 2. The corresponding ZILN parameter estimates for gene expression levels in ZOE-pilot and the IBD data are given in Supplementary Section 1, and the selected sets of parameters in Table 2 cover the distributions well.

Another aspect of this simulation is that the signs of disease effects are randomly perturbed. For example, if every gene is systematically higher expressed—higher nonzero mean values—for the diseased subjects and lower expressed for healthy subjects, despite the significant group mean difference in absolute terms, the mean differences may be negligible in relative terms. In other words, doubling the expressions of every gene in the disease group would not change the composition. This implies that LB may fail to detect signals that have similar magnitudes and are all of the same direction. In reality, only some of the genes might be differentially expressed, while others are not. Also the sign of the difference might vary between genes. Thus, the random perturbation of disease effects signs provides a realistic scenario.

Another interesting scenario is related to possible disease effects on the zero-inflation parameter. When the disease group has systematically higher proportion of zeros than the healthy group, while the nonzero mean values are similar for both groups, the compositional transformation would force the nonzero mean expression of the disease group to be lower than that of the healthy group resulting in false non-zero mean differences between them. This issue, again, can be resolved by the random perturbation of the sign of the effects. In the main results of the simulations, Bernoulli distribution with 50% chance is used for perturbation, and results based on 25% and 0% chances are also presented for LB tests.

No.	δ_μ	δ_θ	δ_π	scenario name
D1	0	0	0	null effect
D2	1	0	0	μ effect
D3	0	1	0	θ effect
D4	0	0	-1	π effect
D5	1	1	0	$\mu\&\theta$ effect I
D6	1	0	-1	$\mu\&\pi$ effect I
D7	0	1	-1	$\theta\&\pi$ effect I
D8	1	0	1	$\mu\&\pi$ effect II
D9	1	-1	0	$\mu\&\theta$ effect II
D10	0	-1	-1	$\theta\&\pi$ effect II

Table 2: Disease effects. Additive effect size on log- (for μ and θ) or logit- (for π) transformed scale; e.g., under $\mu\&\pi$ effect I (D6) without batch effects, the disease (healthy) group has one unit higher (lower) nonzero mean on the log scale and one unit lower (higher) zero proportion on the logit scale than the baseline.

No.	κ_μ	κ_θ	κ_π	scenario name
K1	0	0	0	null effect
K2	0.5	0.5	-0.5	small effect I
K3	1	1	-1	large effect I
K4	0.5	-0.5	-0.5	small effect II
K5	1	-1	-1	large effect II

Table 3: Batch effects. Additive effect size on log- (for μ and θ) or logit- (for π) transformed scale; e.g., under a large effect I (K3), within a disease group, a batch group has 2×1 higher nonzero mean and dispersion on the log scale and 2×1 unit lower zero proportion on the logit scale compared to the other batch group.

4.3.3 Batch effects

For each of the alternative distributions, we further considered batch differences. As batch effects are in most cases nuisance parameters, we considered limited settings and only binary effects were modeled. Batch effects are reflected on parameters for each health and disease group in a similar way as that of disease effects.

Let $\kappa \equiv (\kappa_\mu, \kappa_\theta, \kappa_\pi)$ denote the batch effects such that $\xi_{d,1} = g(\xi_d, \kappa)$ and $\xi_{d,2} = g(\xi_d, -\kappa)$ are the distribution parameters for disease group $d(d = D \text{or} H)$ in batches 1 and 2, respectively. Alternatively, we denote $\xi_{D,1} = g(\xi, \delta, \kappa)$.

The batch effects for the ZILN model are estimated from randomly chosen 300 genes in the ZOE 2.0 dataset assuming that there are no batch effects. These are presented in Figure 2 C. The quartiles of the κ_μ estimates are 0.3, 0.6, and 1.0 in the order. Those for the κ_θ estimates are 0.3, 0.7, and 1.2. Those for the κ_π estimates with finite values are 0.4, 0.8, and 1.4.

Based on the ZILN parameter estimates, we select sets of batch effect parameters for the ZILN model as in Table 3. The parameter estimates for the ZINB are provided in Supplementary Figure 7C. The corresponding ZILN parameter estimates for the gene-species joint data and the species marginal data are given in Supplementary Figure 3C and 5C. The estimated ZILN parameters of the gene expression distribution for genes have a range that overlaps reasonably with Figure 2C and the sets of parameters in Table 3. The corresponding ZILN parameter estimates for gene expression levels in ZOE-pilot and the IBD data are given in Supplementary Section 1, and the selected sets of parameters in Table 3 provide good coverage of the distributions.

5 Results

5.1 Goodness-of-fit of the generative models

5.1.1 Goodness-of-fit in the ZOE data

We use the Lilliefors procedure [57] to evaluate the goodness-of-fit of the generative models (ZILN, ZIG and ZINB) in the ZOE data. The Lilliefors procedure is a data-adaptive version of the Kolmogorov-Smirnov (KS) test [58], where the empirical distribution function (EDF) of a specific gene is compared to the cumulative distribution function (CDF) of the estimated model rather than to a fixed CDF. We randomly select 300 genes and the maximal difference of the EDF and the CDF for each gene is calculated. A p-value is calculated for each gene based on a null distribution generated by Monte Carlo simulations, and the histogram of the p-values and the proportion of the p-values less than the 0.05 threshold are reported. The code for the procedure is available at https://github.com/Hunyong/microbiome2020/blob/master/Readme_KS_test.Rmd.

The Lilliefors procedure is only applied to the non-zero values and ZINB is not evaluated with this procedure. This is because the KS test and the Lilliefors procedure are designed for continuous distributions, while the zero-inflation components in ZILN and ZIG have virtually perfect goodness-of-fit. The Beta distribution, the continuous part of the ZIB model, is also considered in this evaluation. For ZINB distributions, a graphical comparison is provided as an alternative to the quantitative procedure. Estimation of distribution parameters is based on three scaling/transformation methods: RPK, TPM, and arcsine.

Figure 3B suggests that for a small sample, all three generative models appear to have a reasonable fit to the data—the rejection rates are at most 10%. However, the overall high rejection rates in Figure 3A (ZOE 2.0) suggests that the reasonable high rejection rates in Figure 3B (ZOE-pilot) are probably due to typical lower testing power when sample size is small. Despite the high rejection rates in many settings in Figure 3A, the log-normal model shows a consistently good fit. In both Figures 3A and 3B, the TPM normalization is shown to provide a better fit compared to the RPKs, and the arcsine transformation further enhances the rejection rate. It is noteworthy that the results are almost identical between the TPM and the arcsine transformations for the log-normal distribution. This is because, with a large number of genes in the data ($n_{genes} > 300,000$ for ZOE 2.0), most compositional (TPM/c) values are close to zero and, consequently, the arcsine transformation is essentially equivalent the square root transformation with scaling:

$$arcsin(\sqrt{x}) = \sqrt{x} + O(x^{\frac{3}{2}}), \quad \text{as } x \to 0^{+}.$$

This implies that the arcsine transformation is merely a location-shift transformation in the log-normal model for most of compositional values.

The model fit of the ZINB distribution is illustrated in Supplementary Section 3.2 for a couple of randomly chosen genes after rounding values to nearest integers. The results suggest that the ZINB distribution has overall a reasonable fit to the RPK or the TPM transformed data, but has a poor fit to the arcsine transformed data.

5.1.2 Goodness-of-fit of the generative models – the IBD data

Figure 3C shows that the goodness of fit of the IBD data is overall worse than that of the ZOE-pilot data that have a similar sample size. However, for both data, the log-normal distribution with either the TPM or the arcsin transformation yields the best fit.

5.2 Type-I error

Figure 4 presents the type-I error rates of each test for selected representative scenarios. The full results are provided in the Supplementary Material (Supplementary sections 4–6). For MGS, the software does not accommodate batch effects in its current version (1.24.1), and ANCOM-BC also does not control the batch effects. The results of the methods under the batch effects scenarios are still reported. LB, LN, KW, KW-II, and MGS have type-I error rates close to the nominal significance level. The LB has inflated type-I error rates especially when π is large and when the sample size is small ($n = 80$). In contrast, LN, KW, KW-II, and MGS have type-I error rates lower than or equal to the nominal significance level even when π is large in a small sample size.

The type-I error is less stably controlled especially in the two-part models, such as LB and DESeq2-ZINBWaVE, when the zero-inflation (or zero proportion) parameter is high. This is likely a consequence of the nonzero part of those models relying on a small number of nonzero values, that causes high finite sample bias. For example, for a ZILN sample of size $n = 80$, $\pi = 0.9$ means that there are only 8 nonzero values on average and that large sample theory may not be applicable. The inflated or deflated type-I error of those methods dissolves or attenuates when the sample size is large, further implying that finite sample bias is the culprit. Thus, we suggest that, when the sample size is not large and has a high proportion of zeros, the two-part models are not recommended without knowledge that the posited distribution of the test and the true underlying distribution of data match.

MAST, however, frequently has type-I error that is higher than the nominal significance level even for a larger sample size. This becomes even more evident when there are batch effects, e.g., $\kappa = (1, -1, -1)^\top, (0.5, 0.5, -0.5)^\top$, and $(1, 1, -1)^\top$. Hence, MAST needs to be used with caution.

ANCOM-BC2 frequently has an inflated type-I error, which is more apparent under the batch-effects scenarios due to model mis-specification. Its type-I error is often amplified with a larger sample size, implying the existence of systematic bias. ANCOM-BC1, or the original version, has a considerable inflation of type-I error for most of the scenarios, especially under the high zero-proportion scenarios. See Supplementary Figures 9–10. Identification of the structural zeros in ANCOM-BC1 could be unreliable under the high zero-proportion settings.

Finally, DESeq2 has a very low type-I error when used to model these zero-inflated data. Because it was designed for negative binomial distributions without zero-inflation, this may not be a surprising result. On the other hand, DESeq2-ZINBWaVE has on average higher type-I error than DESeq2. However, it has higher-than-nominal type-I errors for larger baseline nonzero mean values, and the inflation becomes even larger for a large sample size, implying that the aberration may not be attributable to the finite sample bias. Designed for the scRNAseq and with its unstable control of type-I error, the ZINB-WAVE extension of DESeq2 should be used with caution.

5.3 Power in a small sample size

The rejection rates at 5% cutoff under alternative distributions, or the power of tests, are illustrated in Figure 5 for selected scenarios of the ZILN model and sample size of $n = 80$. The corresponding results for sample size of $n = 400$ are presented in Figure 6 of Section 5.4. The simulations based on the generative models other than ZILN are discussed in Section 5.5. The full results under all scenarios, i.e., all combinations of generative models, baseline distribution, batch effects, and disease effects, are provided in Supplementary Figure 9 to 14 in Supplementary Sections 4 to 6. We further vary either the significance level or the disease effect size to provide a more comprehensive landscape of the test performances in Sections 5.6 and present an experiment result in Section 5.7 regarding the effect perturbation discussed in Section 4.3.2. In what follows, the power, illustrated in Figure 5, is discussed according to different disease effect scenarios (D2–D8).

D2 $(\mu_D > \mu_H)$. When the disease status is only associated with the difference in nonzero means (μ), LB, MGS, MAST, ANCOM-BC2, and DESeq2-ZINBWaVE have the highest powers for most of the baseline scenarios. However, while the high power of MAST, ANCOM-BC2, and DESeq2-ZINBWaVE comes at a cost of inflated type-I error, the type-I error of LB is relatively reasonably controlled and that of MGS is well controlled. KW-II, that often has one of the highest powers, has relatively weak power for the high zero-proportion scenarios. LN has good power for many baseline scenarios but lacks power when the zero-proportion is high ($\pi = 0.9$). This is due to the bias from model misspecification of the LN model. KW suffers from low power with even smaller π. DESeq2 has a reasonably good power when the zero inflation is not high ($\pi \leq 0.6$). However, it often has lower than 5% power when the data are sparse. This again can be explained by DESeq2's inability to model zero-inflation.

D3 $(\theta_D > \theta_H)$. Most tests lack power in detecting difference in θ (D3), which is expected as all the tests considered in this paper detect the marginal or conditional mean differences and θ difference alone does not affect the mean. However, there are quite a few methods that have power greater than 5%. Methods with inflated type-I error, such as MAST, are expected to have rejection rates higher than 5%. Also, the equal variance assumption that is implied in methods could be the source of the inflation. See, for example, θ_g in (3) of LB and σ_g^2 in (5) of MAST.

D4 $(\pi_D > \pi_H)$. Differences only in π are captured by methods such as LB, LN, KW, and MAST. MAST's slightly higher power than the other methods is counterbalanced by inflated type-I error rates. For these methods, relatively low power for baseline $\pi = 0.9$ and high power for baseline $\pi = 0.6$ can be explained by the design of the experiments. When the baseline π is close to 1 or 0, the absolute difference $(\pi_D - \pi_H)$ between two groups is relatively smaller than that when the baseline π is close to 0.5.

D6 $(\mu_D > \mu_H, \pi_D < \pi_H)$. Rejection rates are higher for D6 $(\mu_D > \mu_H, \pi_D < \pi_H)$ than for both D2 $(\mu_D > \mu_H)$ and D4 $(\pi_D < \pi_H)$, as D6 is expected to have larger marginal mean differences than D2 and D4. As a result, most tests have powers ≥ 0.50 for $\pi \leq 0.9$ including LN and KW.

D8 ($\mu_D > \mu_H, \pi_D > \pi_H$). The disease effect scenario D8 is complicated, as the signal from μ difference and that from π difference offsets the effect on the marginal mean. Thus, tests based on marginal models, i.e. single-part models such as LN, KW, MGS, DESeq2, and ANCOM-BC, inherently cannot avoid low power under this scenario, because they do not separate two opposite signals from two parts. Consequently, they have lower rejection rates under D8 than under either D2 or D4. In contrast, two-part models (LB, MAST, KW-II) entertain the two distinct signals resulting in almost the same power as in D6.

Other scenarios involving θ (D5, D7, D9, D10). Other scenarios involving θ such as D5 ($\mu_D > \mu_H, \theta_D > \theta_H$), D7 ($\theta_D > \theta_H, \pi_D < \pi_H$), D9($\mu_D > \mu_H, \theta_D < \theta_H$), and D10($\theta_D < \theta_H, \pi_D < \pi_H$) do not have remarkable differences in results than the corresponding scenarios without θ effects, i.e., D2, D4, D2, and D4, respectively (Supplementary Figure 2). This is expected, because θ differences do not affect the marginal mean difference and the methods considered in this paper treat θ as a nuisance parameter.

Power in the presence of batch effects. The presence of batch effects affects power even when the batch information is incorporated in the tests. This could be due to the fact that batch effects are made multiplicatively in the generative models, while the models in tests only consider main effects of diseases and batches without their interaction. However, the unevenness of power across different batch-effect scenarios is neither dramatic nor systematic. The patterns, e.g. higher power for MAST and LB, lower power for large π values and so forth, discussed in earlier sections, still hold across different batch-effect scenarios.

Summary. In reality, differential expression only in nonzero mean (D2), only in zero proportion (D4), or in both nonzero mean and zero proportion with the opposite direction (D6), is of most interest and is more feasibly observed than the others. Thus, the LB and MGS tests that have high power under D2 and D6 and the LN test that has high power under D4 and D6 are noteworthy. KW and KW-II tests have high power under D6. However, KW has very low power for most of the settings of D2 and KW-II suffers from low power when $\pi = 0.9$ under all of D2, D4, and D6.

5.4 Power in a larger sample size

As expected, rejection rates are higher in a larger sample size ($n = 400$) and many tests under most scenarios have power close to 1. The patterns for the larger sample size are mostly the same as those under the smaller sample size; MAST, LB, and MGS have the highest power under most scenarios, two-part models have higher power than single-part models when signals are in the opposite directions as in D8, and θ difference (D3) is not properly detected for most of the tests.

5.5 Power under ZINB and ZIG

The patterns of rejection rates under ZINB and ZIG models are not very different from those under ZILN models. The full results are presented in Supplementary Figure 11 to 14.

5.6 Sensitivity analysis of power

To provide a broader view of the power of the tests we vary the significance level and the disease effect size, respectively. Figure 7 presents the power according to the cut-off values ranging from 0 to 0.2 under a few baseline distributions and disease effects scenarios of the ZILN model without batch effects. Each curve either dominates or is dominated by the others for most of the settings uniformly over the cut-off values in $[0.0, 0.2]$, which suggests that the pattern of the previous results is mostly preserved with a different choice of cut-off values.

Figure 8 illustrates the power of the tests under different sizes of disease effects for a subset of the baseline scenarios and $n = 80$ without batch effects. The pattern of lower (higher) power for smaller (larger) disease effect sizes was expected. However, it is noteworthy that when there are only small disease effects on the non-zero mean (i.e., Scenario D11), some methods have virtually no power (DESeq2 and KW) or very low power (LN), suggesting that to compensate for the lack of fit, the effect size should be large enough.

5.7 Power under unbalanced perturbations

As mentioned in Section 4.3.2, LB is based on compositional data where its power attenuates when the disease effects have the same sign across all genes. All the results presented above, i.e., Figures 4 to 6, are based on random perturbation of directions with 50% probability. To illustrate the effect of unbalanced perturbations, we present LB test results under 25% and 0% perturbation, of which rejection rates are presented in Figure 9.

The results indicate that the power of detecting non-zero mean differences diminishes as the probability of perturbed directions diminishes below 50%. However, it should be noted that this subtlety of compositional data only affects the non-zero or Beta model part and not the logistic model part of LB. As discussed in the preceding section, when the disease has a positive effect on zero-proportion uniformly across all genes, false signals are induced on the non-zero mean values by the compositional transformation. This results in inflated power of the LB test under Scenario D4. However, when the perturbation probability is close to 50%, the false positive signals disappear and the power of LB is no more inflated.

6 Application

6.1 Application to the ZOE 2.0 data

We apply the methods that were shown to have reasonable performance with a controlled type-I error in simulations to the ZOE 2.0 data. Because the ZOE 2.0 data have batch effects—significantly different sequencing depths between the two sequencing waves—we do not apply MGS, and hence, LB and LN tests are selected for the analysis. The data are normalized according to the TPM format with an average scale of 20 million. Differences in expression levels in TPM for each gene between health (non-ECC) and disease (ECC) participants are tested after controlling for batch effects and age (coded in months). The data set includes 297 children of ages between 36-71 months (3–5 years old). There are 402,937 genes from 204 bacterial species. Genes with $< 10\%$ prevalence and average TPM < 0.2 were excluded from the analysis, resulting in 157,113 genes in the final analysis data set. For the

LN test, the minimum positive value (0.007) is uniformly added to the TPM values. Because this application is for illustration purposes, we did not apply a multiple testing correction and report only crude p-values.

The p-values of each of the 157,113 genes for the LN and LB tests are summarized in Figure 10. Clearly, the non-uniform shaped histograms for both LB (A) and LN (C) tests in Figure 10 indicate the existence of differentially expressed genes between health and dental disease, at a nominal statistical significance level. The higher peak observed in (A) than (C) implies that there are more genes found to be statistically significant by the LN model than by the LB model. Figure 10 (B) shows that between the two LB model parts, the discrete (logistic) model part yields more significant results than the continuous (beta) model part. This indicates that the significance of the Wald test statistic (C) is mostly driven by the discrete model part. The scatter plot (D) of the coefficients from the two parts and the regression line thereon exhibit only a weak relationship between the two model parts. Genes with Wald statistic p-values are less than 10^{-5} are circled on plot (D) and have same-sign coefficients in both parts, where positive coefficients imply higher nonzero mean (μ) and higher nonzero proportion ($1-\pi$), respectively. However, noting that the coefficients of the continuous model part are close to zero, it appears that the significance is driven by the discrete model part.

This weak relationship may strengthen or weaken the justification for the Wald statistic. In rare events where the signals from the two parts are both strong and with opposite signs, the Wald statistic can detect the signals that would have vanished if the two effects were marginalized. On the other hand, when the signal from only one part is strong, while the other is not, the Wald statistic may not be able to detect the strong signal after being diluted by the weak one, resulting to low power. In this case, using the minimum p-values from both parts with an adjusted significance level, i.e., twice the nominal level for the Bonferroni-type adjustment, could be an alternative strategy.

The number of significantly differentially expressed genes ($p < 10^{-5}$) are summarized in Figure 11. There are more number of significant genes according to the LN test (184) than the LB test (6 for the global test, 30 for the discrete part, and 1 for the continuous part). This is congruent with the fact that the LN test is more powerful than the LB test under D4 (the differential disease effects in zero proportion) in Figure 9. Most of the genes found significant in the LB models are also reported as significant by the LN models. The ten genes with the lowest p-values from the LN models are: C8PIH7, C8PI10, C8PHV7, C8PEV7, C8PKZ2, C8PJY1, C8PG93, C8PKG9, C8PH26, and C8PJD1. Significantly differentially expressed genes according to the LB Wald test were E0DI62, C8PHV7, C8PEV7, C8PI10, C8PIH7, and C8PHV8. The species and proteins associated with those genes and their functions are present in Supplementary Table 10.

The results for the gene-species joint data and for the species marginal data are provided in Supplementary Sections 7 of the Supplementary Material. The patterns are overall similar to those obtained in the gene marginal data analysis; hiked frequency at the low p-value areas for the LN and global LB tests, significance mostly comes from the discrete part than the continuous part, the directions of the two parts in the significant taxonomic units are only weakly consistent with each other. The significant features ($p < 10^{-5}$) in the gene-species joint data analysis using the LB models are E0DJ07 Corynebacterium matruchotii; C8PHV7, C8PHV8, and C8PEV7 Campylobacter gracilis; A3CQN5 Streptococcus cristatus; G1WEB2 Prevotella oulorum; and C7NCB2 Leptotrichia shahii. The ten most significant gene-species from the LN tests were all associated with Campylobacter gracilis and included C8PHV7, C8PHV8, C8PEV7, C8PKG9, C8PI10, C8PH26, C8PHR6, C8PIH7, C8PFD0, and C8PG15.

The significant ($p < 0.01$) taxa from the species marginal data analysis using the LB models include Campylobacter gracilis, Streptococcus cristatus, Leptotrichia hofstadii, Lachnoanaerobaculum saburreum, Leptotrichia shahii, Streptococcus mutans, Campylobacter concisus, Prevotella oulorum. None of the species had $p < 0.01$ in the LN tests.

Streptococcus mutans, one of the identified species, is the most well-document dental caries-associated pathogen. The species most strongly associated with childhood dental caries in this analysis was Campylobacter gracilis, a gram-negative anaerobic bacillus, traditionally isolated from gingival crevices and dental biofilms accumulated close to the gingival margin [59]. Oral campylobacters are enriched in genes for lactate metabolism, which plays an important role in the development and maintenance of acidic conditions in cariogenic biofilms as the predominant glucose-derived product, which is considered to be the main acid involved in caries formation [60]. The capacity of Campylobacter species to produce lactate may be contributing to the development and establishment of early childhood caries, as other microorganisms directly associated to caries disease like Streptococcus sp and Leptotrichia sp, which are benefited by this lactate-rich environment [60, 61]. Chalmers et al. (2015) showed that Campylobacters gracilis is associated with severe early childhood caries at a frequency detection rate of 87.5% [62]. Campylobacters gracilis' active genes shown to have a significant association with ECC were associated with essential steps for: (1) bacterial growth (C8PIH7, encodes for an enzyme that catalyzes the first committed step in fatty acid synthesis) [63]; (2) protein biosynthesis and transport (C8PI10, encodes for an enzyme that catalyzes the attachment of serine to its cognate transfer RNA molecule; C8PKZ2, encodes for the enzyme from biosynthesis of diverse amino acids leading to L-lysine, L-threonine, L-methionine and L-isoleucine; C8PJD1, encodes for an amino acid biosynthesis pathway) [64, 65]; (3) protein transport (C8PG93 encodes for twin-arginine translocation (Tat) pathway, which catalyzes the export of proteins from the cytoplasm across the inner/cytoplasmic membrane.) [66]; (4) DNA replication and transcription (C8PJY1, encodes for key enzymes in the synthesis of nucleoside triphosphates molecular precursors of both DNA and RNA) [67]; (5) biofilm formation or adhesion through gene C8PKG9 (encodes for NFACT-R 1 domain-containing protein) [68] and; (6) energy conservation (C8PHR6 encodes for methylenetetrahydrofolate reductase (MTHFR) of acetogenic bacteria during reduction of carbon dioxide with molecular hydrogen to acetate) [69]. Other genes associated with ECC were A3CQN5 (encodes for Ribosomal RNA small subunit methyltransferase A, which play the role of switch proteins in the ribosome assembly in Streptococcus sanguinis) and C7NCB2, which encodes for a multidrug and toxic compound extrusion (MATE) family of efflux pumps to actively transport of a solute across the membrane in Leptotrichia buccalis [70].

6.2 Application to the IBD data

Next, we apply the LB and LN tests to the IBD data to identify the differentially expressed genes. Out of 1,119,472 genes, 103,966 genes with prevalence rate > 0.1 and mean expression level $> 10^{-8}$ in the relative RPKs were tested. Differences in expression levels in TPM for each gene between control (non-IBD, 26 (25%) participants) and cases (IBD, 78 (75%) participants) groups are tested after controlling for batch effects (binary-coded) and sex. The data set includes 104 patients (52 male and 52 female) who were between 5 and 74 years old at the time of diagnosis (for the cases). For the LN test, the minimum positive value (5×10^{-10}) is uniformly added to the TPM values.

The p-values of each of the 103,966 genes for the LN and LB tests are summarized in Figure 12. Similarly to the ZOE data analysis results, the hike on the left end in both LB (A) and LN (C) indicates the existence of differentially expressed genes. However, in Figure 12 (B), only the continuous part has a conspicuous hike while the discrete part is mostly flat, implying that the signal lies massively in the continuous part, which is also confirmed in the scatter plot (D). It is noteworthy that about 4% of genes have very high discrete model coefficients but are insignificant in two clusters around either 26 and -26 on the y-axis. These are a manifestation of the undesirable feature of Wald statistics called the Hauck-Donner effects, where larger disease effects may not always result in a larger statistic and, as a consequence, may yield lower power [71]. This occurs when a phenotype group has prevalence rate of exactly zero or one, while the other group has prevalence rate away from zero and one. The likelihood-ratio test, permutation tests, Fisher's exact test, regularization, and Bayesian approaches are the alternatives to the Wald test. Among the 523 candidate genes with such prevalence rate pattern, no genes were found significant at the significance level of the nominal P value 10^{-5} by the likelihood-ratio test and the Fisher's exact test of which p-values are presented in Supplementary Figure 21.

The numbers of the signficant genes in LN, LB-continuous, and LB-discrete models are given in Figure 13. The ten most statistically significant genes in the LN model are S3BI82, R5Q3H7, R5PRG3, R5Q1H1, R5QAG2, R5QE55, S3CE88, R5QEQ4, R5PLJ0, and, G2T243 while the top ten genes for the LB model are D4WIY6, Q0TKG5, R6W6W2, D1PDG3, Q17UW4, I9USK4, E2ZM16, R7NP61, B0NN15, and U2ZZD9.

7 Discussion

For the first time, we have provided a comprehensive evaluation of the main analysis methods for differential gene expression of metatranscriptomics data. This simulation study design is inspired by the human oral microbiome sequencing data, to which we investigated the goodness of fit of the generative models after scaling or transformation. The methods were evaluated in terms of control of type-I error and power. The best-performing methods were further used for detecting the differentially expressed genes in the ZOE 2.0 oral metatranscriptomics data and the IBD gut metatranscriptomics data. The microbial genes found significantly associated with ECC were reported and interpreted accordingly, and those having significant association with IBD were also presented.

This simulation study provides a guideline for microbiome researchers in choosing proper DE analysis methods. Our simulation framework could be further applied for

method	type-I error	power	zero model	other
LB	fair	high (D2, D4, D6, D8)	◯	Inflated type-I error for high π and small n in D2.
MGS	good	high (D2, D6)	△	No batch control available.
LN	good	good (D4, D6)	×	Low-powered for high π in D2.
MAST	not controlled	high	◯	Type-I error not controlled.
ANCOM-BC	not controlled	good (D4, D6)	△	Type-I error not controlled.
DESeq2	good	low	×	
DESeq2ZI	unstable	good	△	Type-I error not controlled
KW	good	good (D4, D6)	×	
KW2	good	good (low π)	◯	Low-powered when $\pi \geq 90\%$.

Table 4: Summary of performances. DESeq2ZI is the abbreviation of DESeq2-ZINBWaVE. $\bigcirc$, models zero counts and considers the differential effects on zeros in the tests; $\triangle$, models zero counts but the tests only the marginal or non-zero mean differences; $\times$, does not model zeros.

validating current and future DE methods that were not included in this manuscript. In what follows, we summarize the main findings and discuss the limitations of the simulations.

Which method is the best in general?

In Table 4 we summarize the performance of each method evaluated in this study. The simulation results suggest that for metatranscriptomics data, LB and MGS have good power and good control of type-I error under the scenarios that involve μ-differences (D2 and D6). However, the current version of MGS does not control for batch effects and MGS does not have good power under D4 (π-differences) and LB needs to be used with caution as it may have inflated type-I error for high zero proportion with a small sample size. LN has a fair amount of power when there is non-zero mean difference (for low baseline π values) or zero-proportion difference, or when both differences are present with the opposite directions. Both MAST and ANCOM-BC have considerably high power to detect non-zero mean differences, marginal mean differences, and the combinations of the two under many scenarios, but they do not properly control type-I error. DESeq2 without ZINBWave has both low type-I error and power for metatranscriptomics data, and the one with ZINBWave shows unstable control of type-I error. KW as a nonparametric test, has generally lower power compared to other methods. As simulations are based on parametric generative models, the low power of KW is somewhat expected, but KW can still be considered when the data at hand and the model that other tests assume differ to a great degree.

TPM transformation.

According to our simulations, the log-normal distribution has a decent goodness-of-fit to the ZOE data after TPM transformation. This is consistent with the fact that MGS, which assumes zero-inflated log-normal distribution, is one of the highest powered methods.

Two-part models are beneficial in some cases.

Two-part models have advantages in power over single-part ones when the signals come from two different sources with different directions (D8). Specifically this occurs when the non-zero mean difference and the difference in zero-proportion (or zero-inflation probability for ZINB model) both exist and they have the same sign (or the opposite directions in terms of the marginal mean). Even KW-II, the low-powered test under D2 or D4, often performs better than other single-part models when both signals are present.

Distinct generative models.

Some of the variational factors in this simulation may not have practical implications: e.g., generative models and batch effects. Simulation results showed that there were no substantial differences in power between different generative models, even if each generative model might have different tests as the most powerful test. This might be due to the fact that all the generative distributions considered in this paper, ZILN, ZIG, and ZINB, have similar features such as zero-inflation and left-skewed unimodal distributions in their non-zero parts. Batch effect scenarios (K2 to K5) did not appreciably affect power even when the models were not correctly specified; for instance, the disease and batch effects were misspecified in the sense that independently generated multiplicative effects in μ of disease and batches resulted in interaction effects, whereas, the testing models assumed no interaction effects in this study.

Limitation: distributional assumptions.

This simulation study deals with a sizable number of distinct scenarios. However, it does not cover all possible data generative mechanisms. It is based on a combination of a few parametric generative models and a limited number of parameter sets. For example, the differential expression is based on fixed functions such as log-difference and logit-difference, we did not consider (multiplicative) interactions between disease effects and batch effects, and genes' expressions were generated independently from each other. Each of these issues adds a chance that the simulation results may not plausibly represent the true data distributions in real-life experiments. However, we believe that this simulation results provide useful and practical insights regarding the behavior and performance of each test under certain settings, if the data are not too different from the models considered in this simulation.

Limitation: gene independence assumptions.

It must be acknowledged that we assumed independence between genes. Although in reality, it is likely that some dependency or co-expression of genes is at play and it may affect the significance of potential gene set tests [72] or multiple testing adjustment, such dependency would not affect the differential expression analysis at the individual gene level. Furthermore, all non-collective testing methods and the multiple testing procedures assume independence between genes, where the collective testing methods include MAST, MGS, and DESeq2, and p-values are calculated taking into account the dependence between genes using empirical Bayes. In other words, for each gene, the p-values obtained by individual gene tests on independently simulated data are valid; and so are the corresponding type-I and type-II error rates. Although, in real data settings, genes are dependent to some degree, we expect that, the high dimensionality of the data used here (i.e., a large number of simulated genes) likely introduced spurious correlations between features [73], our simulation results may not be far from a realistic scenario. Future work could explore more realistic settings where gene expression levels are correlated.

Funding

This work was supported by grants from the National Institutes of Health, National Institute of Dental and Craniofacial Research, R03-DE028983 and U01-DE025046.

Availability of data and materials

For reproducible research, the code for all simulations is available via https://github.com/Hunyong/microbiome2020, the ZOE 2.0 data are being deposited in the dbGaP repository https://www.ncbi.nlm.nih.gov/gap under the umbrella study name Trans-Omics for Precision Dentistry and Early Childhood Caries or TOPDECC (accession: phs002232.v1.p1), and the IBD data are available at https://ibdmdb.org.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Study design: H.C., B.T., D.W.; Simulations: H.C., B.T., C.L, M.L, D.W; Data analyses: H.C., C.L, B.L.; Bioinformatics: J.R., A.R., D.W.; Manuscript: H.C., C.L., B.T., A.R., K.D., D.W.; Principal Investigator: D.W.

Author details

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, US. [2]Department of Statistics, University of Connecticut, Storrs, US. [3]Research Computing, University of North Carolina, Chapel Hill, US. [4]Division of Diagnostic Sciences, University of North Carolina, Chapel Hill, US. [5]Department of Genetics, University of North Carolina, Chapel Hill, US. [6]Division of Pediatric and Public Health, University of North Carolina, Chapel Hill, US. [7]Department of Epidemiology, University of North Carolina, Chapel Hill, US. [8]Division of Oral and Craniofacial Health Sciences, University of North Carolina, Chapel Hill, US. [9]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, US.

References

1. Kaakoush, N.O., Day, A.S., Huinao, K.D., Leach, S.T., Lemberg, D.A., Dowd, S.E., Mitchell, H.M.: Microbial dysbiosis in pediatric patients with crohn's disease. Journal of clinical microbiology **50**(10), 3258–3266 (2012)
2. Tilg, H., Kaser, A., *et al.*: Gut microbiome, obesity, and metabolic dysfunction. The Journal of clinical investigation **121**(6), 2126–2132 (2011)
3. Kilian, M., Chapple, I., Hannig, M., Marsh, P., Meuric, V., Pedersen, A., Tonetti, M., Wade, W., Zaura, E.: The oral microbiome–an update for oral healthcare professionals. British dental journal **221**(10), 657–666 (2016)
4. Gopalakrishnan, V., Helmink, B.A., Spencer, C.N., Reuben, A., Wargo, J.A.: The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. Cancer cell **33**(4), 570–580 (2018)
5. Visconti, A., Le Roy, C.I., Rosa, F., Rossi, N., Martin, T.C., Mohney, R.P., Li, W., de Rinaldis, E., Bell, J.T., Venter, J.C., *et al.*: Interplay between the human gut microbiome and host metabolism. Nature communications **10**(1), 1–10 (2019)
6. Olsen, I., Yamazaki, K.: Can oral bacteria affect the microbiome of the gut? Journal of oral microbiology **11**(1), 1586422 (2019)
7. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., *et al.*: Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature **569**(7758), 655–662 (2019)
8. Peterson, S.N., Meissner, T., Su, A.I., Snesrud, E., Ong, A.C., Schork, N.J., Bretz, W.A.: Functional expression of dental plaque microbiota. Frontiers in Cellular and infection microbiology **4**, 108 (2014)
9. Duran-Pinedo, A.E., Chen, T., Teles, R., Starr, J.R., Wang, X., Krishnan, K., Frias-Lopez, J.: Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. The ISME journal **8**(8), 1659–1672 (2014)
10. Duran-Pinedo, A.E.: Metatranscriptomic analyses of the oral microbiome. Periodontology 2000 **85**(1), 28–45 (2021)
11. Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., *et al.*: Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome **5**(1), 27 (2017)
12. Holmes, I., Harris, K., Quince, C.: Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS one **7**(2), 30126 (2012)
13. Willis, A.D.: Rarefaction, alpha diversity, and statistics. Frontiers in microbiology **10**, 2407 (2019)
14. Chen, E.Z., Li, H.: A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics **32**(17), 2611–2617 (2016)
15. Calgaro, M., Romualdi, C., Waldron, L.D., Risso, D., Vitulo, N.: Assessment of statistical methods from single cell, bulk rna-seq and metagenomics applied to microbiome data. bioRxiv (2020)
16. Martin, B.D., Witten, D., Willis, A.D., *et al.*: Modeling microbial abundances and dysbiosis with beta-binomial regression. Annals of Applied Statistics **14**(1), 94–115 (2020)
17. Mallick, H., Ma, S., Franzosa, E.A., Vatanen, T., Morgan, X.C., Huttenhower, C.: Experimental design and quantitative analysis of microbial community multiomics. Genome biology **18**(1), 228 (2017)
18. Niu, S.-Y., Yang, J., McDermaid, A., Zhao, J., Kang, Y., Ma, Q.: Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. Briefings in bioinformatics **19**(6), 1415–1429 (2018)
19. Narayanasamy, S., Jarosz, Y., Muller, E.E., Heintz-Buschart, A., Herold, M., Kaysen, A., Laczny, C.C., Pinel, N., May, P., Wilmes, P.: Imp: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome biology **17**(1), 1–21 (2016)
20. Westreich, S.T., Treiber, M.L., Mills, D.A., Korf, I., Lemay, D.G.: Samsa2: a standalone metatranscriptome analysis pipeline. BMC bioinformatics **19**(1), 1–11 (2018)

21. Hickl, O., Heintz-Buschart, A., Trautwein-Schult, A., Hercog, R., Bork, P., Wilmes, P., Becher, D.: Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut microbiome. Microorganisms **7**(9), 367 (2019)
22. Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nature methods **10**(12), 1200 (2013)
23. McDavid, A., Finak, G., Yajima, M.: MAST: Model-based Analysis of Single Cell Transcriptomics. (2019). R package version 1.8.2. https://github.com/RGLab/MAST/
24. Peng, X., Li, G., Liu, Z.: Zero-inflated beta regression for differential abundance analysis with metagenomics data. Journal of Computational Biology **23**(2), 102–110 (2016)
25. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**(12), 550 (2014)
26. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease **26**(1), 27663 (2015)
27. Divaris, K., Shungin, D., Rodríguez-Cortés, A., Basta, P.V., Roach, J., Cho, H., Wu, D., Zandona, A.G.F., Ginnis, J., Ramamoorthy, S., *et al.*: The supragingival biofilm in early childhood caries: Clinical and laboratory protocols and bioinformatics pipelines supporting metagenomics, metatranscriptomics, and metabolomics studies of the oral microbiome. In: Odontogenesis, pp. 525–548. Springer, ??? (2019)
28. Divaris, K., Slade, G.D., Zandona, A.G.F., Preisser, J.S., Ginnis, J., Simancas-Pallares, M.A., Agler, C.S., Shrestha, P., Karhade, D.S., Ribeiro, A.d.A., Cho, H., Gu, B.Y., Meyer, B.D., Joshi, A.R., Azcarate-Peril, M.A., Basta, P.V., Wu, D., North, K.E.: Cohort profile: Zoe 2.0—a community-based, genetic epidemiologic study of early childhood oral health. International Journal of Environmental Research and Public Health **17**(21), 8056 (2020)
29. Pitts, N.B., Baez, R.J., Diaz-Guillory, C., Donly, K.J., Feldens, C.A., McGrath, C., Phantumvanit, P., Seow, W.K., Sharkov, N., Songpaisan, Y., *et al.*: Early childhood caries: Iapd bangkok declaration. Journal of dentistry for children (Chicago, Ill.) **86**(2), 72 (2019)
30. Lin, H., Peddada, S.D.: Analysis of compositions of microbiomes with bias correction. Nature communications **11**(1), 1–11 (2020)
31. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C.: Metagenomic biomarker discovery and explanation. Genome biology **12**(6), 1–18 (2011)
32. Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., Knight, R.: Establishing microbial composition measurement standards with reference frames. Nature communications **10**(1), 1–11 (2019)
33. Anderson, M.J.: A new method for non-parametric multivariate analysis of variance. Austral ecology **26**(1), 32–46 (2001)
34. Zhao, N., Chen, J., Carroll, I.M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J.J., Ringel, Y., Li, H., Wu, M.C.: Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. The American Journal of Human Genetics **96**(5), 797–807 (2015)
35. Wu, C., Chen, J., Kim, J., Pan, W.: An adaptive association test for microbiome data. Genome medicine **8**(1), 56 (2016)
36. Hu, Y.-J., Satten, G.A.: Testing hypotheses about the microbiome using the linear decomposition model (ldm). Bioinformatics (2020)
37. King, G., Zeng, L.: Logistic regression in rare events data. Political analysis **9**(2), 137–163 (2001)
38. Stasinopoulos, D.M., Rigby, R.A.: Generalized additive models for location scale and shape (gamlss) in r. Journal of Statistical Software **23**(7), 1–46 (2007)
39. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., *et al.*: Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology **16**(1), 278 (2015)
40. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Nature Precedings, 1–1 (2010)
41. Van den Berge, K., Perraudeau, F., Soneson, C., Love, M.I., Risso, D., Vert, J.-P., Robinson, M.D., Dudoit, S., Clement, L.: Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. Genome biology **19**(1), 24 (2018)
42. Paulson, J.N.: Normalization and differential abundance analysis of metagenomic biomarker-gene surveys. PhD thesis, University of Maryland, College Park (2015)
43. Paulson, J.N., Olson, N.D., Braccia, D.J., Wagner, J., Talukder, H., Pop, M., Bravo, H.C.: metagenomeSeq: Statistical analysis for sparse high-throughput sequncing. Bioconductor package. http://www.cbcb.umd.edu/software/metagenomeSeq (2013)
44. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American statistical Association **47**(260), 583–621 (1952)
45. Hothorn, T., Hornik, K., Van De Wiel, M.A., Zeileis, A.: A lego system for conditional inference. The American Statistician **60**(3), 257–263 (2006)
46. Hothorn, T.: Package 'coin' (2019)
47. Lachenbruch, P.A.: Analysis of data with clumping at zero. Biometrische Zeitschrift **18**(5), 351–356 (1976)
48. Taylor, S., Pollard, K.: Hypothesis tests for point-mass mixture data with application to omics data with many zero values. Statistical Applications in Genetics and Molecular Biology **8**(1), 1–43 (2009)
49. Wagner, B.D., Robertson, C.E., Harris, J.K.: Application of two-part statistics for comparison of sequence variant counts. PloS one **6**(5), 20296 (2011)
50. Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., *et al.*: Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol **8**(6), 1002358 (2012)
51. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)

52. McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol **10**(4), 1003531 (2014)
53. Preisser, J.S., Stamm, J.W., Long, D.L., Kincade, M.E.: Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries research **46**(4), 413–423 (2012)
54. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.-P.: A general and flexible method for signal extraction from single-cell rna-seq data. Nature communications **9**(1), 284 (2018)
55. Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., Ballman, K.: An omnibus test for differential distribution analysis of microbiome sequencing data. Bioinformatics **34**(4), 643–651 (2017)
56. Xia, Y., Sun, J., Chen, D.-G.: Statistical Analysis of Microbiome Data with R. Springer, ??? (2018)
57. Lilliefors, H.W.: On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American statistical Association **62**(318), 399–402 (1967)
58. Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. The annals of mathematical statistics **19**(2), 279–281 (1948)
59. Vandamme, P., Daneshvar, M., Dewhirst, F., Paster, B., Kersters, K., Goossens, H., Moss, C.: Chemotaxonomic analyses of bacteroides gracilis and bacteroides ureolyticus and reclassification of b. gracilis as campylobacter gracilis comb. nov. International Journal of Systematic and Evolutionary Microbiology **45**(1), 145–152 (1995)
60. Iraola, G., Perez, R., Naya, H., Paolicchi, F., Pastor, E., Valenzuela, S., Calleros, L., Velilla, A., Hernández, M., Morsella, C.: Genomic evidence for the emergence and evolution of pathogenicity and niche preferences in the genus campylobacter. Genome biology and evolution **6**(9), 2392–2405 (2014)
61. McLean, J.S., Fansler, S.J., Majors, P.D., McAteer, K., Allen, L.Z., Shirtliff, M.E., Lux, R., Shi, W.: Identifying low ph active and lactate-utilizing taxa within oral microbiome communities from healthy children using stable isotope probing techniques. PloS one **7**(3), 32219 (2012)
62. Chalmers, N.I., Oh, K., Hughes, C.V., Pradhan, N., Kanasi, E., Ehrlich, Y., Dewhirst, F.E., Tanner, A.C.: Pulp and plaque microbiotas of children with severe early childhood caries. Journal of oral microbiology **7**(1), 25951 (2015)
63. Freiberg, C., Pohlmann, J., Nell, P., Endermann, R., Schuhmacher, J., Newton, B., Otteneder, M., Lampe, T., Häbich, D., Ziegelbauer, K.: Novel bacterial acetyl coenzyme a carboxylase inhibitors with antibiotic efficacy in vivo. Antimicrobial agents and chemotherapy **50**(8), 2707–2712 (2006)
64. Cox, R.J., Gibson, J.S., Mayo Martín, M.B.: Aspartyl phosphonates and phosphoramidates: The first synthetic inhibitors of bacterial aspartate-semialdehyde dehydrogenase. ChemBioChem **3**(9), 874–886 (2002)
65. Tadrowski, S., Pedroso, M.M., Sieber, V., Larrabee, J.A., Guddat, L.W., Schenk, G.: Metal ions play an essential catalytic role in the mechanism of ketol–acid reductoisomerase. Chemistry–A European Journal **22**(22), 7427–7436 (2016)
66. Stephenson, K.: Sec-dependent protein translocation across biological membranes: evolutionary conservation of an essential protein transport pathway. Molecular membrane biology **22**(1-2), 17–28 (2005)
67. Chargaff, E.: The Nucleic Acids. Elsevier, ??? (2012)
68. Burroughs, A.M., Aravind, L.: A highly conserved family of domains related to the dna-glycosylase fold helps predict multiple novel pathways for rna modifications. RNA biology **11**(4), 360–372 (2014)
69. Bertsch, J., Öppinger, C., Hess, V., Langer, J.D., Müller, V.: Heterotrimeric nadh-oxidizing methylenetetrahydrofolate reductase from the acetogenic bacterium acetobacterium woodii. Journal of bacteriology **197**(9), 1681–1689 (2015)
70. Saier, M.H.: A functional-phylogenetic classification system for transmembrane solute transporters. Microbiology and molecular biology reviews **64**(2), 354–411 (2000)
71. Hauck Jr, W.W., Donner, A.: Wald's test as applied to hypotheses in logit analysis. Journal of the american statistical association **72**(360a), 851–853 (1977)
72. Wu, D., Smyth, G.K.: Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic acids research **40**(17), 133–133 (2012)
73. Fan, J., Han, F., Liu, H.: Challenges of big data analysis. National science review **1**(2), 293–314 (2014)

Figures

Tables

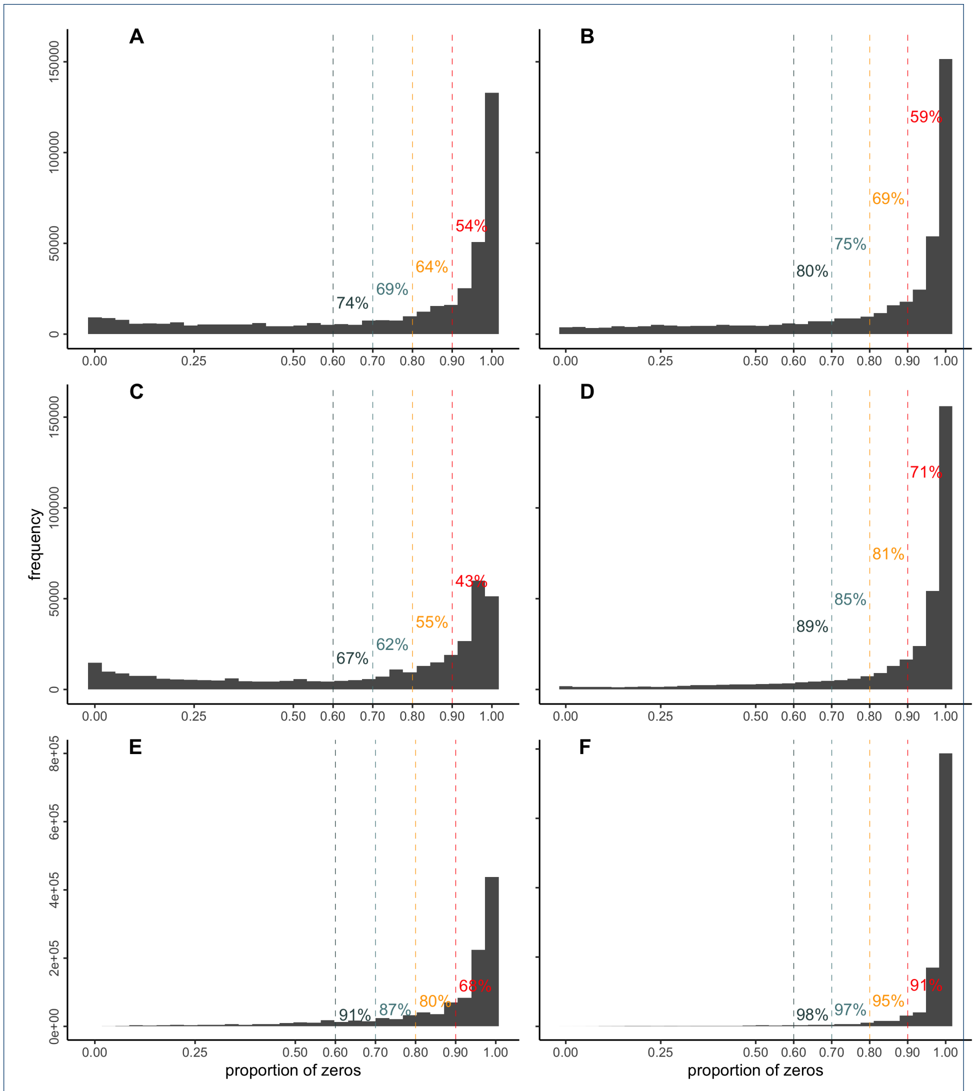

Figure 1: Histogram of zero-proportions at the gene-level in the metagenomics (LEFT) and the metatranscriptomics (RIGHT) data of the ZOE 2.0 (Row 1), the ZOE-pilot (Row 2), and the IBD (Row 3) studies. A. ZOE 2.0 metagenomics; B. ZOE 2.0 metatranscriptomics; C. ZOE-pilot metagenomics; D. ZOE-pilot metatranscriptomics; E. IBD metagenomics; F. IBD metatranscriptomics; Numbers on the histogram represent the proportion of genes of which the zero proportion is greater than or equal to the cutoff values, or the vertical bars left to the numbers.

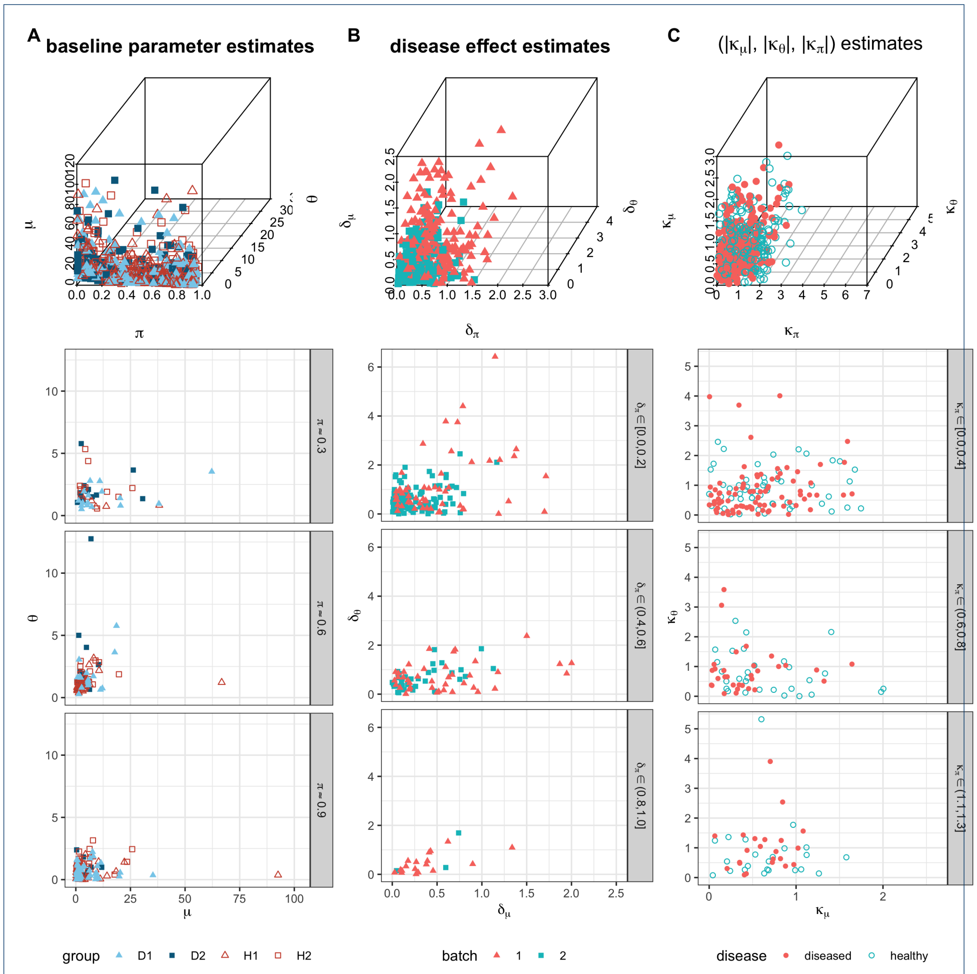

Figure 2: Column A: Parameter estimates of baseline ZILN distributions obtained from the ZOE 2.0 data with the 3-dimensional scatter plot on the top row and each of the subsequent rows representing π estimates being within 0.03 from 0.9, 0.6, and 0.3.

Column B: Disease effect estimates based on ZILN models obtained from the ZOE 2.0 data in absolute values $(|\delta_\mu|, |\delta_\theta|, |\delta_\pi|)$

Column C: Batch effect estimates based on ZILN models obtained from the ZOE 2.0 data in absolute values $(|\kappa_\mu|, |\kappa_\theta|, |\kappa_\pi|)$

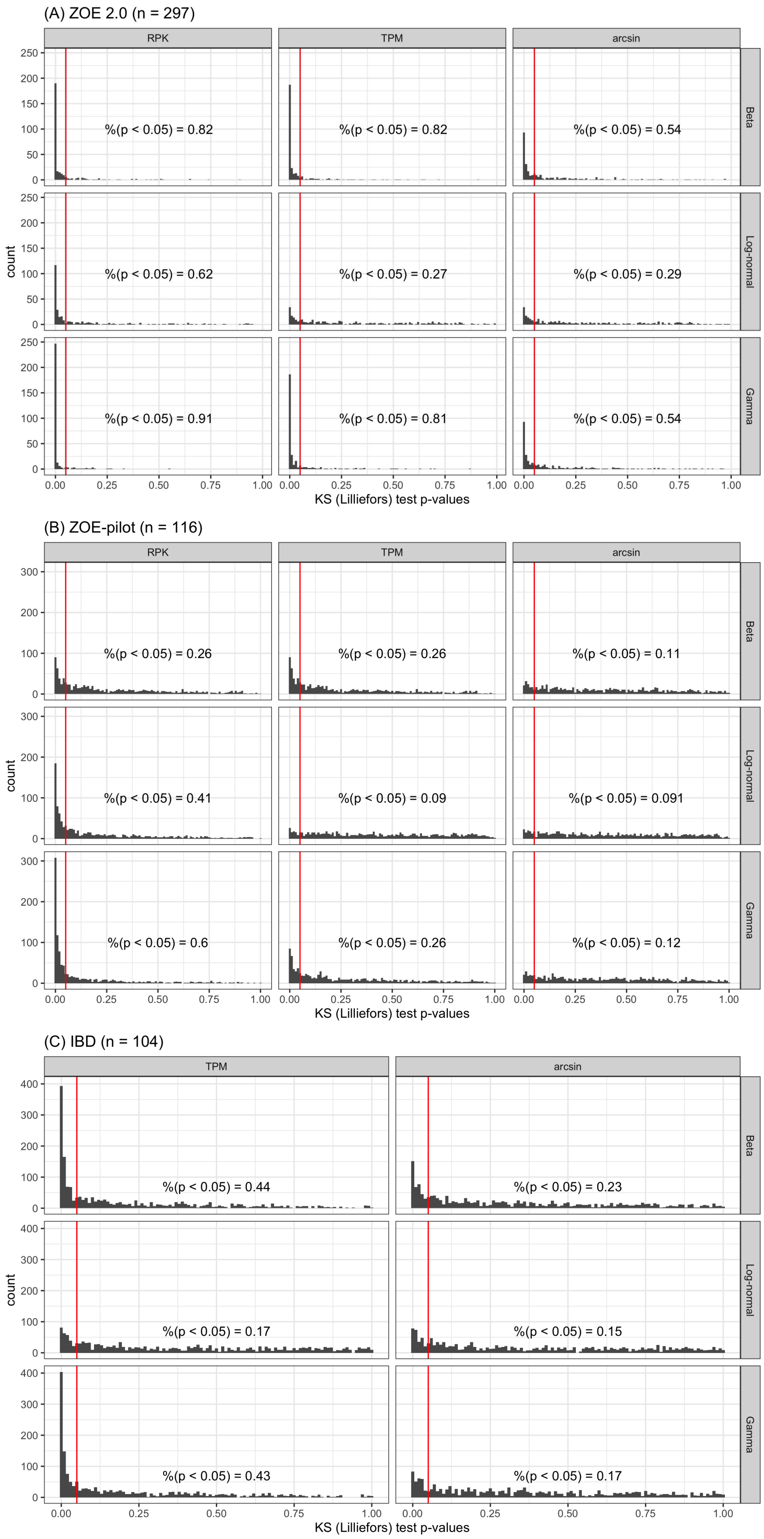

Figure 3: Goodness of fit (Kolmogorov-Smirnov) test results for Beta, Log-normal, and Gamma distributions (rows) with different scaling/transformation methods (columns). The top nine histograms (A) are based on the ZOE 2.0 data ($n = 297$), the middle nine graphs (B) are based on the ZOE-pilot data ($n = 116$), and the bottom six graphs (C) are based on the IBD data ($n = 104$).

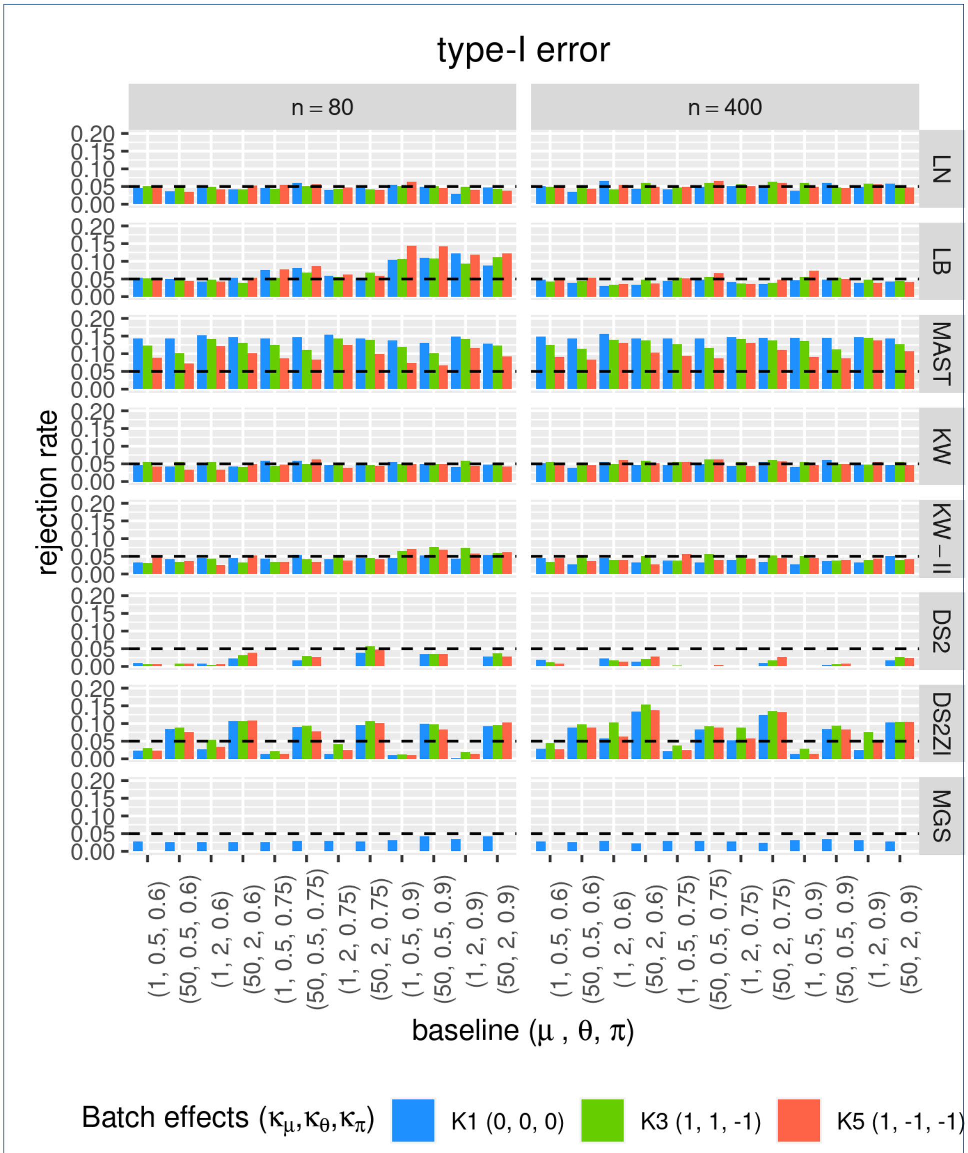

Figure 4: Type-I error rates for ZILN models. Columns and rows correspond to sample sizes and tests, respectively, the $X-$axis represents baseline distributions, and colors represent batch effects. The dotted horizontal lines denote the significance level (5%). DS2 = DESeq2, DS2ZI = DESeq2-ZINBWaVE, ANCOM = ANCOM-BC2.

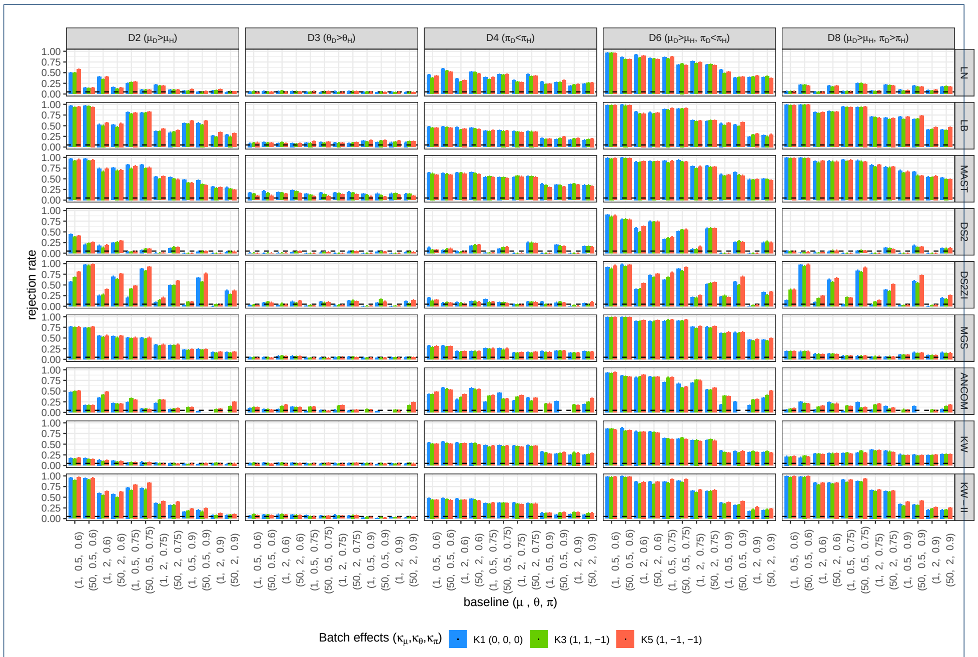

Figure 5: Power under alternative ZILN distributions for a small sample size (n = 80). Columns and rows correspond to tests and alternative distributions, respectively, the $X-$axis represents baseline distributions, and colors represent batch effects. The dotted horizontal lines denote the nominal significance level (5%). DS2 = DESeq2, DS2ZI = DESeq2-ZINBWaVE, ANCOM = ANCOM-BC2.

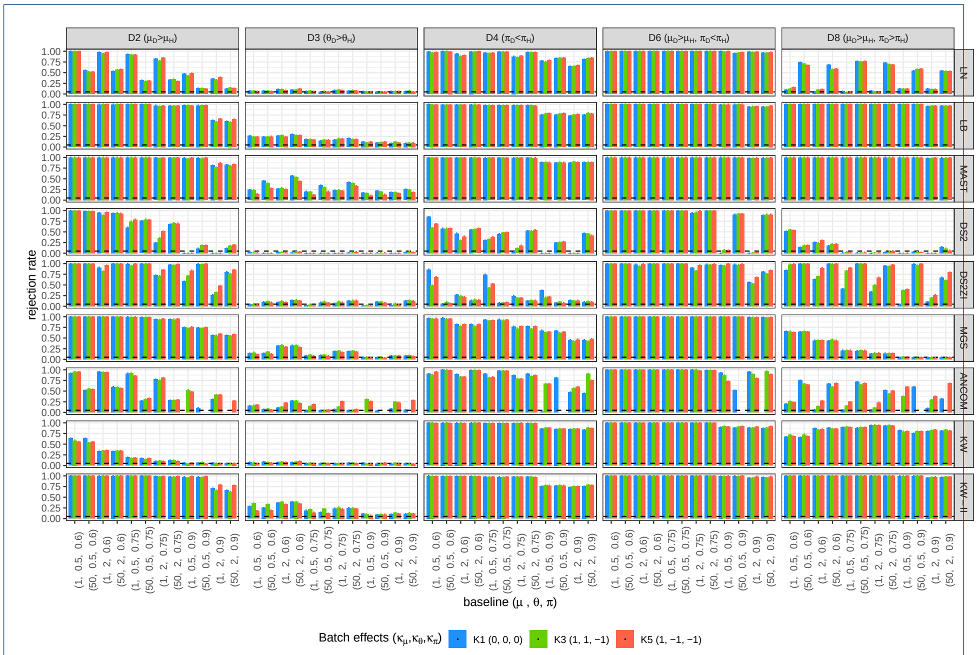

Figure 6: Power under alternative ZILN distributions for a large sample size ($n = 400$). Columns and rows correspond to tests and alternative distributions, respectively, the $X-$axis represents baseline distributions, and colors represent batch effects. The dotted horizontal lines denote the nominal significance level (5%). DS2 = DESeq2, DS2ZI = DESeq2-ZINBWaVE, ANCOM = ANCOM-BC2.

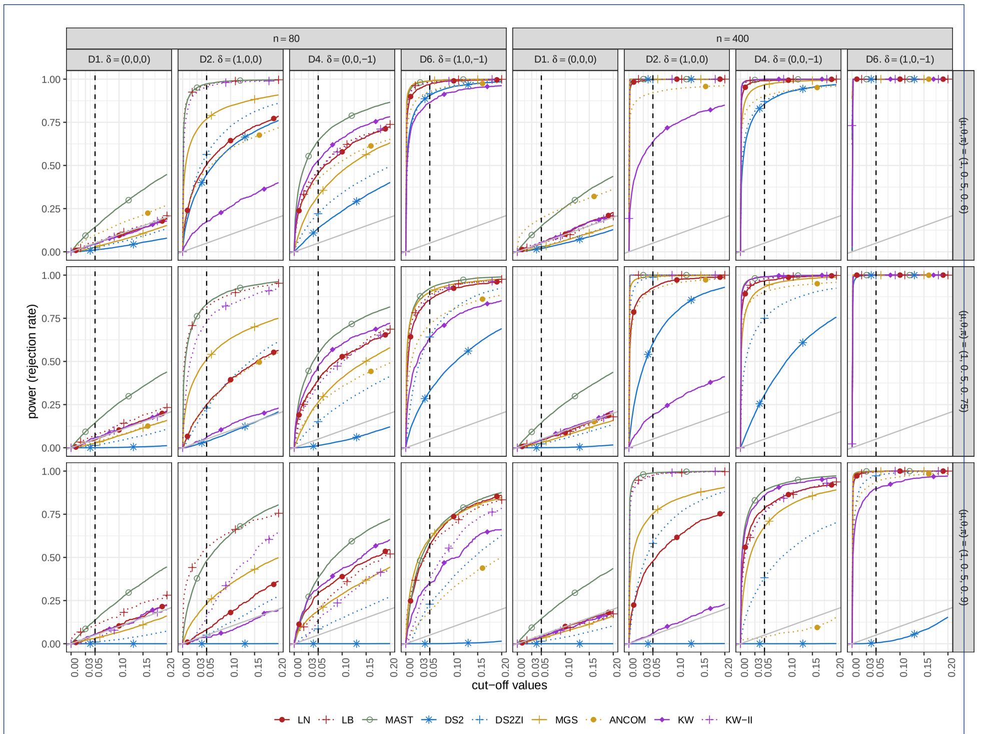

Figure 7: Power curves of differential expression tests according to different cut-off values. No batch effects are assumed. The gray solid diagonal lines denote the nominal significance level.

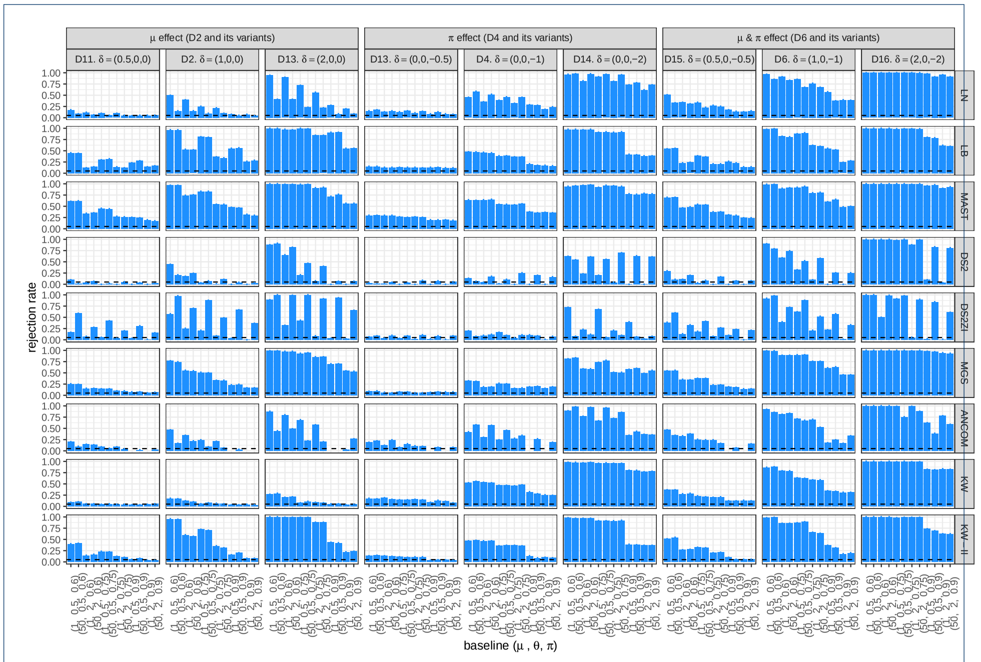

Figure 8: The powers of the differential expression tests according to different effect sizes. No batch effects are assumed.

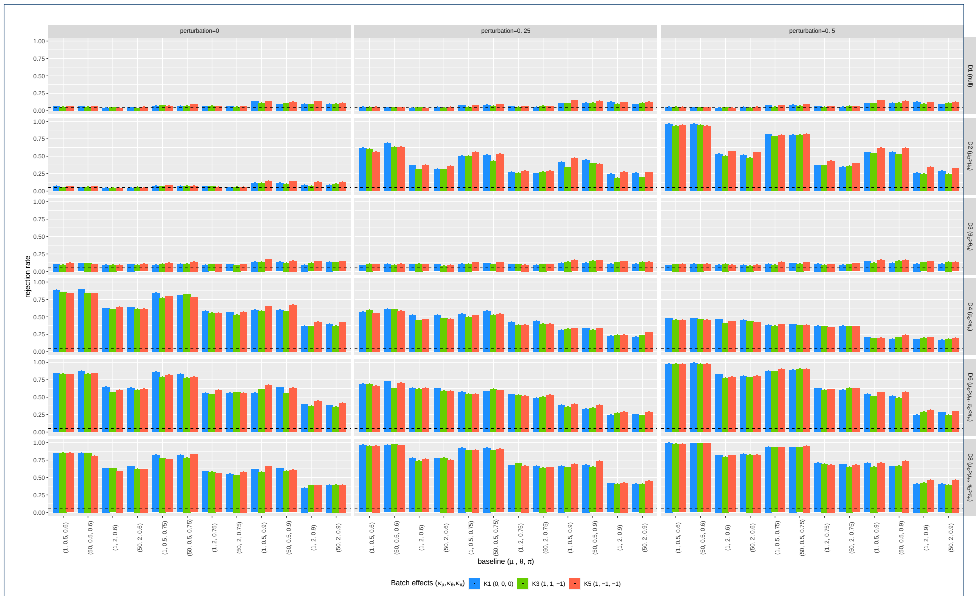

Figure 9: Type-I error rates and power of LB under different probabilities of perturbation.

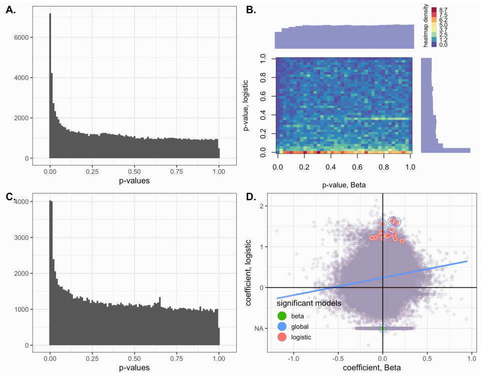

Figure 10: Application to the ZOE 2.0 data analysis results. A. Histogram of the p-values of the log-normal models. B. Histogram of the joint p-values of the logistic Beta models (logistic and Beta parts). C. Histogram of the single p-values of the logistic Beta models (Wald test statistics). D. Scatter plot of the coefficients of the LB models, with the circled dots representing the most significant genes—Wald test statistic $p < 10^{-5}$. NA on the y-axis indicates that the logistic part was not estimated.

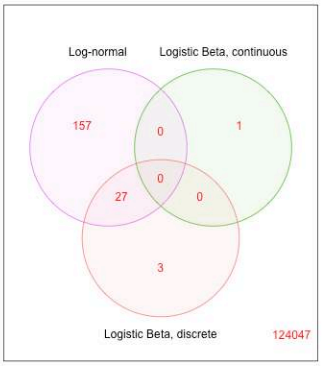

Figure 11: Venn diagram of genes with p-values are less than 10^{-5} for each evaluated model in the ZOE 2.0 data.

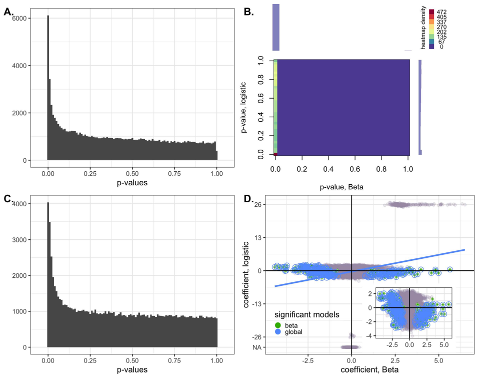

Figure 12: Application to the IBD data analysis results. A. Histogram of the p-values of the log-normal models. B. Histogram of the joint p-values of the logistic Beta models (logistic and Beta parts). C. Histogram of the single p-values of the logistic Beta models (Wald test statistics). D. Scatter plot of the coefficients of the LB models, with the circled dots representing the most significant genes—Wald test statistic $p < 10^{-5}$. NA on the y-axis indicates that the logistic part was not estimated.

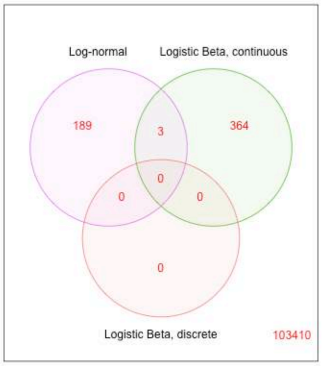

Figure 13: Venn diagram of genes with p-values are less than 10^{-5} for each evaluated model in the IBD data.