Received: 22 July 2019 | Revised: 30 November 2019 | Accepted: 16 December 2019

DOI: 10.1111/1755-0998.13128

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data

Joshua G. Harrison | W. John Calder | Vivaswat Shastry | C. Alex Buerkle

Department of Botany, University of Wyoming, Laramie, WY, USA

Correspondence
Joshua G. Harrison, Department of Botany, 3165, University of Wyoming, 1000 E. University Avenue, Laramie, WY 82071, USA.
Email: joshua.harrison@uwyo.edu

Funding information
National Science Foundation, Grant/Award Number: EPS-1655726; University of Wyoming

Abstract

Molecular ecology regularly requires the analysis of count data that reflect the relative abundance of features of a composition (e.g., taxa in a community, gene transcripts in a tissue). The sampling process that generates these data can be modelled using the multinomial distribution. Replicate multinomial samples inform the relative abundances of features in an underlying Dirichlet distribution. These distributions together form a hierarchical model for relative abundances among replicates and sampling groups. This type of Dirichlet-multinomial modelling (DMM) has been described previously, but its benefits and limitations are largely untested. With simulated data, we quantified the ability of DMM to detect differences in proportions between treatment and control groups, and compared the efficacy of three computational methods to implement DMM—Hamiltonian Monte Carlo (HMC), variational inference (VI), and Gibbs Markov chain Monte Carlo. We report that DMM was better able to detect shifts in relative abundances than analogous analytical tools, while identifying an acceptably low number of false positives. Among methods for implementing DMM, HMC provided the most accurate estimates of relative abundances, and VI was the most computationally efficient. The sensitivity of DMM was exemplified through analysis of previously published data describing lung microbiomes. We report that DMM identified several potentially pathogenic, bacterial taxa as more abundant in the lungs of children who aspirated foreign material during swallowing; these differences went undetected with different statistical approaches. Our results suggest that DMM has strong potential as a statistical method to guide inference in molecular ecology.

KEYWORDS

Bayesian statistics, compositional data analysis, Dirichlet, Hamiltonian Monte Carlo, hierarchical modelling, JAGS, Markov chain Monte Carlo, microbial ecology, microbiome, multinomial, stan, transcriptome, variational inference

1 | INTRODUCTION

In many scientific disciplines, data from both manipulative experiments and surveys of natural variation are often counts of observations that are assigned to categories. Given some total level of observational effort, the counts of the different features in the sample (e.g., taxa or transcripts) reflect the underlying proportions of those features in the sampled composition (e.g., an assemblage of organisms or collection of molecules). In molecular ecology, such sampling can take the form of detecting and counting taxa based on observed DNA sequences (e.g., in molecular barcoding or microbial ecology) or counting the reads assigned to specific transcripts in studies of gene expression (Fernandes et al., 2014; Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017; Tsilimigras & Fodor, 2016). For these applications, sampling effort corresponds to the total number of sequence reads, and the count of reads assigned to a taxon or gene supports inference of their true proportion in the composition. Moreover, the total number of reads that can be obtained is constrained by the sequencing instrument, with reads ascribed to samples and features within each sample. Due to this constant sum constraint, compositional data have the important quality that as the relative abundance of one feature in the composition increases, other features must decrease.

Molecular ecologists often rely on compositional count data to define differences between sampling groups. As an example, we may wish to know how the foliar and root microbiomes of a particular plant taxon differ. To answer this question, an understanding of how each feature shifts in relative abundance among sampling groups is required. In our view, if even a single feature shifts in relative abundance among groups, then this demonstrates an effect of sampling group that could be biologically interesting, albeit subtle. Such effects will go unnoticed if analyses rely on techniques such as ordination and PERMANOVA, which can provide insight into overall differences between sampling groups (McKnight et al., 2019), but provide no statistical model to identify those features that may differ in relative abundance among groups. Accordingly, a variety of methods have been developed to perform the seemingly simple task of determining treatment-induced shifts in relative abundance, which is often referred to as 'differential relative abundance testing' or 'differential expression' testing (the latter phrase arises because the roots of many of these methods lie within the field of functional genomics; Bullard, Purdom, Hansen, & Dudoit, 2010; Dillies et al., 2013; Paulson, Stine, Bravo, & Pop, 2013; Thorsen et al., 2016; Weiss et al., 2017).

Methods for detecting shifts in relative abundance vary tremendously—and the benefits and drawbacks of various methods are the subjects of an ongoing dialogue (Bullard et al., 2010; McMurdie & Holmes, 2014; Weiss et al., 2017). Early approaches typically relied on repeated frequentist tests after transforming count data to account for differences in sampling effort among replicates or sampling groups, typically via rarefaction, conversion to proportions, or, for transcriptomic data, reads per kilobase per million mapped reads (Bullard et al., 2010). More recently, rarefaction has been criticized because it can amplify the variation present within replicates and thus reduce statistical power (McMurdie & Holmes, 2014; but see McKnight et al., 2019 and Weiss et al., 2017 for counterarguments). Numerous statistical modelling approaches have arisen to account for the challenges imposed by compositional data, while avoiding rarefaction. These methods often model feature relative abundance and typically involve some form of normalization followed by repeated frequentist testing. Methods most often differ in the choice of distribution(s) utilized for modelling and normalization method employed. For example, the software DESEQ2 (Love, Huber, & Anders, 2014) and EDGER (Robinson, McCarthy, & Smyth, 2010) are widely-used for analysis of gene expression data and, more recently, for microbiome analysis (Weiss et al., 2017). These tools model feature relative abundances using a negative binomial distribution (a reparameterization of the Poisson distribution to allow for overdispersion), which is scaled to account for variation in sequencing depth among samples (each tool uses different normalization methods). Next a generalized linear model is used to determine if features differ in relative abundance between sampling groups. By comparison, the popular ANCOM software applies a centred log ratio transformation (Aitchison, 1982) to the data followed by repeated parametric or nonparametric testing (depending on the data) with multiple comparison correction. These few examples serve to illustrate the variety of approaches available for performing differential expression testing. However, we are unaware of any popular method that allows estimates of feature relative abundance to be easily extracted while preserving the uncertainty in those estimates for propagation to downstream analyses. This perceived need led us to consider modelling feature relative abundances using the Dirichlet and multinomial distributions (Box 1) in a Bayesian framework.

The multinomial and Dirichlet probability distributions are the relevant models of the aforementioned sampling process that commonly leads to compositional data (Figure 1). Statistical modelling using these distributions has proven successful in a number of biological studies. For instance, Fordyce, Gompert, Forister, and Nice (2011) rely on Dirichlet-multinomial modelling (DMM) to analyze ecological count data, such as counts of behavioural and dietary choices of animals (also see Coblentz, Rosenblatt, & Novak, 2017). Similar models have been applied to large counts of DNA sequences—for instance, Fernandes et al. (2014; ALDEX2), Nowicka and Robinson (2016; DRIM-SEQ), and Rosa et al. (2012; HMP) use DMM to estimate and compare feature-specific relative abundances in transcriptomes and microbiomes. Additionally, DMM has been used to model mixtures of compositions, a situation that could arise in a laboratory-derived microbial assemblage occurring as a contaminant within samples, or in mixtures of different communities in nature (MICROBEDMM, Holmes, Harris, & Quince, 2012; SOURCETRACKER, Knights et al., 2011; BIOMICO, Shafiei et al., 2015; FEAST, Shenhav et al., 2019; ECOSTRUCTURE, White, Dey, Mohan, Stephens, & Price, 2019). Likewise, DMM has been used to estimate association networks among microbial taxa (SPARCC, Friedman & Alm, 2012; MLDM, Yang, Chen, & Chen, 2017).

These models represent important advances and demonstrate the utility of DMM, but it remains unclear how data attributes, such as rank-abundance profiles and dimensionality, affect the accuracy and precision of parameter estimates. Moreover, compared to models that rely on other distributions or are based on different statistical methods (likelihood and frequentist methods), Bayesian DMM can be computationally demanding. Recent advances in computational statistics such as Hamiltonian Monte Carlo (HMC) sampling and variational inference (VI, see Section 2; Blei, Kucukelbir, & McAuliffe, 2017; Monnahan, Thorson, & Branch, 2017) may improve model runtime, but the accuracy and performance of these new methods remains to be evaluated in different modelling contexts.

Consequently, we conducted a simulation experiment to learn the limits and benefits of DMM through the analysis of data that encompass much of the variety in attributes encountered across scientific domains (e.g. replication, number of observations, and so on; Figure 2). Notably, included in simulated data, were those emulating the results of high-throughput sequencing of microbial assemblages, as these are analytically challenging due to their dimensionality, high among-replicate variation, and extreme rank-abundance skew—often several microbial taxa are orders of magnitude more abundant than the numerous marginal taxa that typically compose the bulk of biodiversity within a sample (e.g., see Lynch & Neufeld, 2015; Sachdeva, Campbell, & Heidelberg, 2019). Our primary analytical goal was to measure the sensitivity and accuracy of DMM for comparing feature relative abundance between compositions and to compare the performance of DMM with competing approaches. Also, we provide a primer on the requisite algorithmic methods (e.g., VI and HMC) for Bayesian implementation of DMM and explore how different algorithms affect model accuracy and computational expense. Finally, we analyzed a data set published by Duvallet et al. (2019) that describes the lung microbiomes of children experiencing aspiration of foreign material and evaluated to what extent DMM recapitulated the published analyses or detected additional differences among microbiomes.

2 | METHODS

2.1 | Dirichlet multinomial modelling approach

Our specification of the Dirichlet-multinomial model generally follows that of Fordyce et al. (2011; implemented in the BAYESPREF software) and takes as input a matrix of counts (X). The rows of this matrix correspond to different replicates ($\vec{x}_i$; the superscripted arrow denotes a vector) and the columns correspond to features of the composition (the format of an OTU or transcript table). Each count x_{ij} in this matrix corresponds to the jth feature (of n features in total) in the composition observed in the ith replicate sample. Replicates are grouped into k groups, corresponding to treatment conditions, sampling locations, or some other stratification that specifies which replicates share information (parameters shared among replicates for the group). Counts in each row of the matrix are multinomially distributed:

$$\vec{x}_i \sim \text{Multinomial}\left(\vec{p}_i, N_i\right).$$

Each value p_{ij} in $\vec{p}_i$ is the probability of observing a particular feature j in sample i and $\vec{N}$ is a vector of the total counts in each sample. The product across i replicates of the i multinomial distributions forms the likelihood in the model and can be written:

$$P\left(\vec{x}_{1\ldots i} | \vec{p}_{1\ldots i}, N_{1\ldots i}\right) = \prod_i \frac{N_i!}{x_{i1}! \cdots x_{ij}!} p_{i1}^{x_{i1}} \cdots p_{ij}^{x_{ij}}$$

The prior probability for the vector of feature proportions ($\vec{p}_i$) is a Dirichlet distribution, with parameters that are specific to the kth group of replicates and that are learned from the data:

$$\vec{p}_i \sim \text{Dirichlet}\left(\vec{\pi}_k \theta_k\right).$$

$$\theta_k \sim \text{Uniform}\left(0{,}4000\right).$$

In this parameterization of the Dirichlet distribution for $\vec{p}_i$, the $\vec{\pi}_k$ parameters correspond to the expected proportions of each of the n features (e.g., a particular transcript or taxon) in group k, and θ is an intensity parameter that is shared among all features (see Box 1). For a given $\vec{\pi}$, larger θ means less variation among deviates from the Dirichlet expectation $\vec{\pi}$. The probability density function of this distribution, across i replicates within the kth group, is given by,

$$P(\vec{p}_i | \vec{\pi}, \theta) = \frac{1}{B\left(\vec{\pi}\theta\right)} \prod_j p_{ij}^{\pi_j \theta - 1}$$

$$B\left(\vec{\pi}\theta\right) = \frac{\prod_j \Gamma\left(\vec{\pi}_j \theta\right)}{\Gamma\left(\sum_j \vec{\pi}_j \theta\right)}$$

where $B\left(\vec{\pi}\theta\right)$ is a normalizing function that ensures the Dirichlet distribution integrates to one. The hyperprior for the $\vec{\pi}_k$ parameters at the 'topmost', or most inclusive, level of the model hierarchy is another Dirichlet distribution with equal prior probability for each feature within the composition. For this Dirichlet distribution we use $\alpha_{1\ldots n} = 10^{-7}$ as a prior that will contribute little information, gives an expected value of $1/n$, and has a high variance on the expectation:

$$\vec{\pi}_k \sim \text{Dirichlet}\left(\vec{\alpha}\right).$$

The overall model for the posterior distribution for parameters of a sampling group is:

$$P(\vec{p}_{1\ldots i}, \vec{\pi}, \vec{\alpha}, \theta | X, \vec{N}) \propto \left(\prod_i P(\vec{x}_i | \vec{p}_i, N_i) P(\vec{p}_i | \vec{\pi}, \theta)\right) P(\vec{\pi} | \vec{\alpha}) P\left(\vec{\alpha}\right) P\left(\theta\right).$$

To quantify differences in proportions of features between two sampling groups [often referred to as 'differential relative abundance testing'; Thorsen et al., 2016, Weiss et al., 2017), posterior probability distributions (PPDs) for $\vec{\pi}_{j,k=1} - \vec{\pi}_{j,k=2}$ (Figure 2d) can be obtained. Consistent with convention, if 95% of the samples of this PPD of differences are either greater or less than zero, then there is a high certainty of a nonzero effect of sampling group on feature relative abundance. One can also observe where zero occurs in the PPD of differences to quantify the probability of no effect of sampling group on feature relative abundance.

If a sampling scheme was used that induces dependence among replicates via a more nested hierarchical structure then the model described above, then the model hierarchy could be extended to include inference of the Dirichlet distributions describing the relative abundances of features within each additional stratum of the

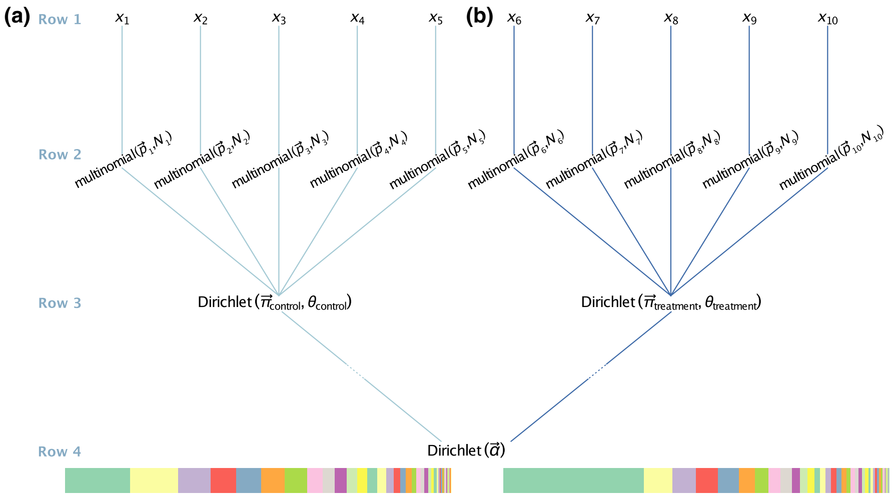

(a)
Row 1
x_1
x_2
x_3
x_4
x_5
(b)
x_6
x_7
x_8
x_9
x_{10}
Row 2
multinomial($\vec{p}_1$,N_1)
multinomial($\vec{p}_2$,N_2)
multinomial($\vec{p}_3$,N_3)
multinomial($\vec{p}_4$,N_4)
multinomial($\vec{p}_5$,N_5)
multinomial($\vec{p}_6$,N_6)
multinomial($\vec{p}_7$,N_7)
multinomial($\vec{p}_8$,N_8)
multinomial($\vec{p}_9$,N_9)
multinomial($\vec{p}_{10}$,N_{10})
Row 3
Dirichlet ($\vec{\pi}_{\text{control}}$, θ_{control})
Dirichlet ($\vec{\pi}_{\text{treatment}}$, $\theta_{\text{treatment}}$)
Row 4
Dirichlet ($\vec{\alpha}$)

FIGURE 1 Visual depiction of hierarchical Bayesian modelling of the relative abundance of features within compositional data. Panels (a) and (b) represent two different sampling groups–a treatment group and a control group. The coloured bars at the bottom of the plot show two hypothetical compositions that differ between those sampling groups. These compositions differ by virtue of the first feature, shown in pastel green, shifting dramatically in relative abundance, thus all other features are shifted in relative abundance as well (because these are proportion data and must sum to one). This interdependency represents an opportunity for statistical modelling because parameters that describe relative abundance are mutually informative. However, interdependency also poses many challenges (see main text). Replicates within sampling groups are denoted as x_i, where i is in an integer in the range [1,10] (row 1). Replicates consist of data that are multinomially distributed (see Box 1). Therefore, each replicate is modelled using a unique multinomial distribution with parameters $\vec{p}_i$ and N_i (row 2), where the vector $\vec{p}$ describes the probabilities that an observation would be assigned to a particular feature and the N parameters denote the total number of observations per replicate. Multinomial parameters are modelled as a deviate from a Dirichlet distribution unique to the treatment group (row 3). The $\vec{\pi}$ parameters of the Dirichlet are estimates of proportional abundance for all features within the group. The θ parameter is a scalar intensity parameter that describes the amount of among-replicate variation present within each sampling group. The prior imposed on the Dirichlet distributions of both sampling groups has the expectation $1/n$ for each feature, where n is the number of features. If desired, additional Dirichlet distributions could be added between rows three and four to share information as dictated by more complexly nested experimental designs [Colour figure can be viewed at wileyonlinelibrary.com]

sampling scheme. For example, consider a study design where subjects are provided one of several diets and gut microbiome samples are taken from both sexes. In this case, one would want to account for non-independence among the data due to both sex and diet treatment. This can be accomplished through incorporation of additional Dirichlet distributions into the model, $P(\vec{\pi}_k|\vec{\psi}_m,\tau)$, where $\vec{\psi}_m$ describes the relative abundances of features within each diet treatment (m), τ is the intensity parameter for that Dirichlet distribution, and $\vec{\pi}_k$ describes relative abundances of features within each sex that is nested within each diet treatment. In this way, the model can be extended to encompass as many hierarchical layers as desired, given suitable sampling and replication (Coblentz et al., 2017).

2.1.1 | A primer of the algorithms to perform DMM

One goal of statistical modelling is to estimate values for parameters that could correspond with directly observable variables (i.e., the data) or with latent, unobservable, variables (i.e., those that are inferred from observable variables). Bayesian modelling attempts to estimate parameters of interest, while explicitly quantifying the uncertainty in those estimates and allowing for the influence of prior knowledge on estimates. Much of Bayesian statistical modelling relies on Markov chain Monte Carlo (MCMC) sampling (Gelman et al., 2013). A Markov chain is a series of states where each state depends upon the immediately preceding state. Monte Carlo refers to repeated, random sampling. MCMC is a process by which values are suggested randomly from a probability distribution and substituted into the functions that define the model. Over MCMC iterations, sampling converges on the most supported parameter space (the PPDs for model parameters) and samples in the chain occur with probability defined by the PPD.

There are several MCMC algorithms and they primarily differ in how they choose or propose new values and their criteria for inclusion of those values in the chain (Gelman et al., 2013). A standard MCMC tool is the Metropolis algorithm (Gelman et al., 2013, page 289). To perform Metropolis sampling, a value (x_t) is proposed from some distribution $Q(x_t|x_{t-1})$, where t is iteration (a suitable initial value, x_0, is required). Once x_t is chosen a ratio of $\alpha = \frac{f(x_t)}{f(x_{t-1})}$ is calculated, where $f(x)$ is a function that is proportional to the probability density to be estimated. The new value x_t is accepted into the chain with probability α, otherwise $x_t = x_{t-1}$. The Metropolis algorithm relies on a symmetric proposal distribution, such that $Q(x_t|x_{t-1}) = Q(x_{t-1}|x_t)$. The Metropolis-Hastings (MH) algorithm extends this concept through relaxing the assumption of symmetry regarding the proposal probability distribution.

Gibbs sampling (Geman & Geman, 1987; Kruschke, 2015) is a special case of the MH algorithm (because the proposal acceptance

Box 1 A brief explanation of the multinomial and Dirichlet distributions

The multinomial distribution is the multivariate generalization of the binomial distribution. The binomial distribution can be used to describe counts of binary outcomes, with respective probabilities p and $1-p$. For instance, with a finite sample of observations, the binomial distribution would be useful for estimating the frequency of females (p) in a dioecious population. The multinomial distribution extends this concept to encompass more than two unique outcomes. For instance, a composition comprising three equally abundant features would have the the following multinomial parameter vector: $\vec{p}=\left[\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$. As an example, consider data from a sequencing machine. The counts of sequences that fall into each category (e.g., transcripts or taxa) are multinomially distributed, with a probability that corresponds to its relative abundance. For three equally abundant features (i.e., microbial taxa), there would be an equal chance of sampling a sequence from each of the features and on average we would expect to obtain the same number of sequences from each (for this example, we assume no laboratory-technique imposed bias).

To share information among samples in the same sampling group (e.g., treatment group, host population, or sampling location) and recover group-level estimates of the proportion of each feature in a composition, the Dirichlet distribution can be appropriately parameterized. The Dirichlet distribution is the multivariate generalization of the beta distribution. Deviates from a standard beta distribution fall in the range of $[0,1]$, and the distribution can be parameterized with expectation π (the expected frequency of the reference category, with $1-\pi$ for the alternative category) and a parameter, θ, that affects the variation among deviates. Likewise, the Dirichlet distribution can be parameterized by a vector of expected frequencies of each feature ($\vec{\pi}$), and an intensity parameter, θ. When drawing deviates from the Dirichlet distribution, the intensity parameter influences the amount of among-deviate variation in the frequencies observed—for a given $\vec{\pi}$, larger intensity parameters induce less among-deviate variation. This parameterization of the Dirichlet thus allows modelling of the variation among experimental replicates (the 'noise' within the data).

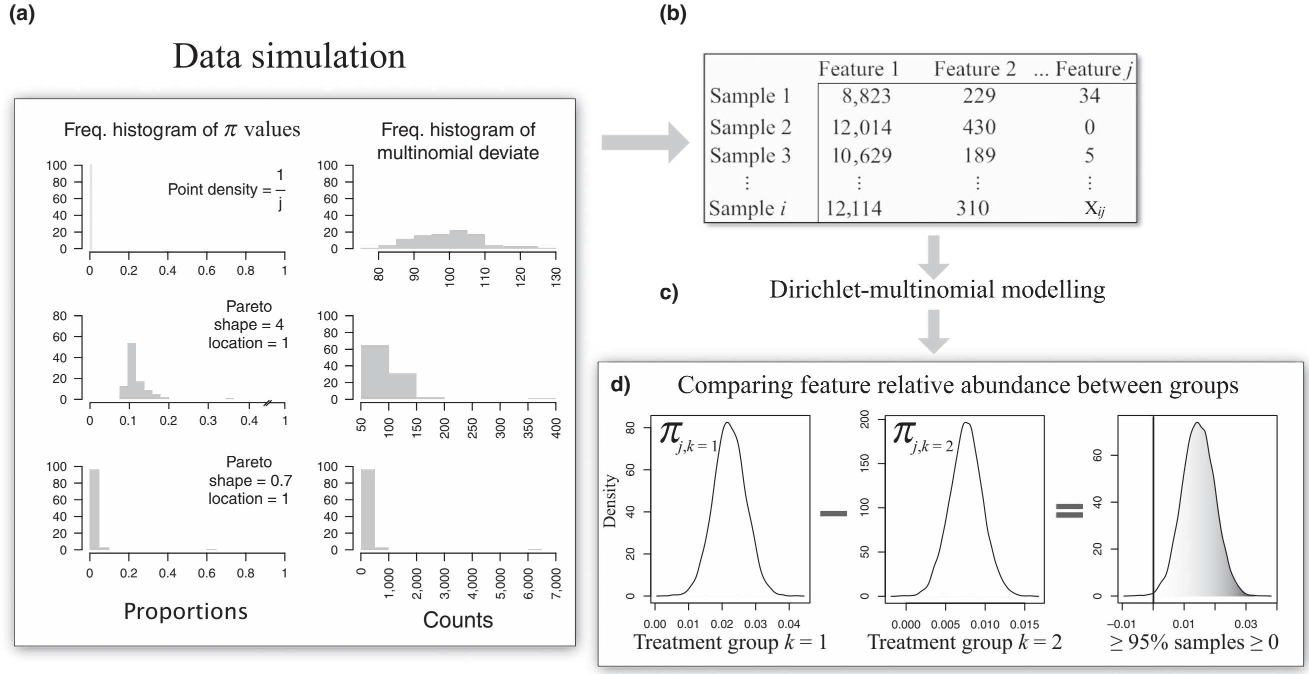
Information about the frequencies of features within replicates ($\vec{p}$) is shared to estimate frequencies for each feature within that sampling group ($\vec{\pi}$), forming a hierarchical model (Figure 1) that is analogous to how replicates can be used in an analysis of variance to learn about marginal, grand means associated with treatments. Estimates of frequencies of compositional features at the sampling group level ($\vec{\pi}$) are the basis of inferences about which features differ among sampling groups (e.g., treatment vs. control) and by how much (on an absolute or normalized scale).

criterion is always met; see page 289 in Gelman et al., 2013) and is suited for cases when the distributions used within the model are conditionally conjugate, such as when the prior and likelihood distributions are conjugate and, consequently, their product has a well defined form. At each iteration of Gibbs sampling (t), each parameter is sampled from the conditional distribution defined by the other parameters in the model, which are held constant at values chosen at iteration $t - 1$. Parameters are typically updated one at a time, in a predefined order.

The probabilistic programming language JAGS (Plummer, 2003) implements Gibbs and Metropolis-Hastings MCMC as required by a specified model structure. Henceforth, we refer to parameter estimation via Gibbs, Metropolis, and Metropolis-Hasting sampling as MCMC. These algorithms can be slow to converge for complex models; indeed in our experience, in a JAGS implementation, convergence may not be observed for the majority of parameters over a week of runtime for DMM with high dimensional data (such as transcriptomic data), even with sensible chain initialization values (a bespoke software implementation of MCMC tuned to the data and model would likely be faster, but would require greater care in programming and use).

Hamiltonian Monte Carlo seeks to improve upon the efficiency of MCMC through the use of a physics inspired algorithm (for an excellent description of HMC see Monnahan et al., 2017). The sampling method can be envisioned by considering a ball being dropped into a bowl and allowed the ball to roll about the curvature of the bowl. The bowl is the PPD and is frictionless, so the ball will roll back and forth in the bowl forever. After repeated drops of the ball into the bowl, from different angles and with different potential energies, the shape of the PPD is determined from the combined paths the ball took across all iterations. The benefit of this approach is that samples from nearly anywhere in the PPD can be generated at each iteration (HMC does not use a Markov chain process, but does rely on a Metropolis ratio to determine acceptability of updates), whereas MCMC typically chooses values based on the previous state space and thus cannot quickly move throughout the PPD, which can slow chain mixing and time to convergence. The probabilistic programming language and software STAN allows the use of an improved version of HMC called the 'no U-turn' sampler that avoids redundant sampling of parameter space (Hoffman & Gelman, 2014). To continue the previous analogy, when the ball starts to make a U-turn due to the curvature of the bowl, the sampler is stopped, and the ball dropped again—thus avoiding spending sampler time in previously explored parameter space.

Hamiltonian Monte Carlo often improves model runtime (Monnahan et al., 2017) over MCMC, but can still be quite time consuming. Variational inference is a class of optimization methods from the machine learning literature that can rapidly approximate PPDs (Blei et al., 2017), and thus holds great promise for statistical modelling of complex data where the speed of MCMC or HMC is insufficient. Variational inference (VI) has yet to be widely applied by biologists, but it has been used to estimate population genetic structure (Raj, Stephens, & Pritchard, 2014; Scordato et al., 2017),


(a)
Data simulation
Freq. histogram of π values
Freq. histogram of multinomial deviate
Point density = 1/j
Pareto shape = 4 location = 1
Pareto shape = 0.7 location = 1
Proportions
Counts
(b)
c)
Dirichlet-multinomial modelling
d)
Comparing feature relative abundance between groups
Density
$\pi_{j,k=1}$
$\pi_{j,k=2}$
Treatment group $k = 1$
Treatment group $k = 2$
≥ 95% samples ≥ 0

	Feature 1	Feature 2	... Feature *j*
Sample 1	8,823	229	34
Sample 2	12,014	430	0
Sample 3	10,629	189	5
⋮	⋮	⋮	⋮
Sample *i*	12,114	310	X_{ij}

FIGURE 2 Visual description of simulation approach. (a) Deviates from either of two Pareto distributions, or a point density, defined as one divided by the number of features (*j*), were used to simulate values used to parameterize Dirichlet distributions. The use of these three approaches generated deviates with parameters that differed dramatically in rank abundance profiles, as shown in the left portion of panel (a). These deviates were, in turn, used to parameterize a Dirichlet distribution (with intensity parameter θ). A deviate of this Dirichlet distribution served as the parameter vector of a multinomial distribution that was sampled (b) to generate a feature (*j*) by replicate (*i*) matrix that emulated an OTU or transcript table (see the right portion of panel a for an example frequency distribution of multinomial deviates). This matrix encompassed samples belonging to two sampling groups. Dirichlet parameters for each group were made to differ such that certain features varied in relative abundance between groups by a known effect size. (c) Hierarchical Bayesian modelling (Figure 1) was used to estimate the Dirichlet parameters ($\pi_{j,k}$) describing the relative abundance of each feature (*j*) in each sampling group (*k*). (d) To determine if a feature (π_j) differed in relative abundance between treatment groups (*k*), the posterior probability distribution (PPD) for the feature of interest from one treatment group, $\pi_{j,k=1}$, was subtracted from the PPD for that feature from the second treatment group, $\pi_{j,k=2}$. If the resulting PPD of differences indicated zero difference was improbable, then there was high certainty that π_j differed between treatment groups. Additionally, the location of zero within the PPD quantified the certainty of a nonzero effect of treatment

genotype-phenotype associations (Carbonetto & Stephens, 2012; Logsdon, Hoffman, & Mezey, 2010), phylogenetic relationships (Jojic et al., 2004), and in a generalized latent linear modelling context (Niku et al., 2019).

The idea behind VI is that the exact PPD need not be estimated, but can be approximated through optimization of parameters of more tractable distributions. Briefly, a density is chosen from a family of distributions and optimized so that the Kullback-Leibler (KL) divergence between that density and the PPD is minimized. KL divergence relies on the definition of entropy. Entropy is a measure of the information present within a distribution and can be expressed (for a discrete probability distribution):

$$H = -\sum_{i=1}^{N} p\left(x_i\right) \log p\left(x_i\right).$$

where $p(x)$ is a function that outputs a probability contingent upon an input value x, which is indexed by i. It is perhaps easiest to intuit entropy using $\log_2$, in which case H is the minimum number of bits needed to encode the data under consideration. KL divergence extends this idea to quantify the amount of information necessary to explain the divergence (||) between two probability distributions p and q, which, in this example, are discrete:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p\left(x_i\right)\left(\log p\left(x_i\right) - \log q\left(x_i\right)\right).$$

Because this measure of divergence is based on the quantification of entropy, when p and q differ greatly, then more information is required to explain how they differ and KL divergence increases. For VI we wish to minimize the KL divergence between the probability distribution $p(\vec{z}|\vec{x})$ and some density $q\left(\vec{z}\right)$ chosen from a family of distributions Q. To avoid computation of $p\left(\vec{x}\right)$ (see Blei et al., 2017, for more), minimizing the KL divergence can be solved by maximizing the 'evidence lower bound' (ELBO; the $\mathbb{E}$ used below refers to expectation):

$$\text{ELBO}(q) = \mathbb{E}\left[\log p\left(\vec{z},\vec{x}\right)\right] - \mathbb{E}\left[\log q\left(\vec{z}\right)\right].$$

The ELBO is the negative of KL divergence after adding the constant $\log p\,(x)$. Thus maximizing the ELBO is equivalent to minimizing the KL divergence, up to the added constant. This also means that:

$$\log p\left(\vec{x}\right) = D_{\mathrm{KL}}(p||q) + \mathrm{ELBO}\left(q\right).$$

The ELBO describes the lower bound of the evidence, because when the ELBO is subtracted from the evidence ($\log p\,(x)$) the result must be ≥ 0, because KL must be ≥ 0. Because maximizing the ELBO does not require computing $\log p\left(\vec{x}\right)$ it is easier than minimizing KL divergence. Maximization techniques can then be used to find the density $q^{*}\left(\vec{z}\right)$ that best approximates $p(\vec{z}|\vec{x})$.

Choosing Q such that the family of densities includes a $q^{*}\left(\vec{z}\right)$ that provides a good approximation, while being easily optimized, is the challenge of VI. STAN solves this problem through a method called 'automatic differentiation variational inference' (Kucukelbir, Ranganath, Gelman, & Blei, 2015) by first transforming the data that are the support of the latent variables to lie within the real numbers ($\mathbb{R}$) and then suggesting a Gaussian distribution, which can be optimized to fit the data, and which induces a non-Gaussian approximation to the untransformed data. STAN's default approach uses the 'mean-field' algorithm, which treats latent variables (z_j) as independent and assigns a unique density, $q_j\left(z_j\right)$, to each of these j variables. Since STAN transforms the data such that latent variables have support on $\mathbb{R}$ and then fits Gaussian distributions to those data, this statement becomes the product of many Gaussian distributions, each of which are optimized to minimize the ELBO. Following the notation of Blei et al. (2017), this can be written:

$$q\left(\vec{z}\right) = \prod_{j=1}^{m} q_j\left(z_j\right).$$

VI is an attractive technique because it can be many orders of magnitude faster than MCMC (Raj et al., 2014).). However, it is unclear how well VI works across analytical tasks and model specifications (Blei et al., 2017).

2.2 | Model implementation

We performed DMM in the R statistical computing environment (R Core Team, 2019) using models specified for the JAGS and STAN (Carpenter et al., 2017) software programs, and used the models through the RJAGS (Plummer, 2015) and RSTAN (Stan Development Team, 2018) R packages, respectively. JAGS uses MCMC (Gibbs and MH), whereas STAN implements HMC (no U-turn sampling) and VI. Model specification for use in STAN was slightly modified from that described above in that we used an exponential distribution as the form of the prior for θ_k:

$$\theta_k \sim \mathrm{Exponential}\left(\lambda = 0.001\right).$$

This change in model specification followed the recommendation to avoid uniform priors provided in the STAN documentation.

For HMC and MCMC implementations of DMM, we used two chains to explore parameter space. Initial values for $\vec{p}_i$ in each chain were the vector of proportions observed from the data in replicate i, and values for $\vec{\pi}_k$ were initialized using the vector of observed proportions for each feature across replicates within k (i.e., the maximum likelihood estimates for $\vec{p}_i$ and $\vec{\pi}_k$). θ was left to be initialized internally by RJAGS and RSTAN. In RJAGS, the model was subjected to an adaptation period long enough for the sampler to approach optimal efficiency as determined via internal heuristics, or for 20,000 iterations, whichever came first. Models were updated ('burned in') for 300,000 steps for RJAGS and 1,000 steps for RSTAN (with a maximum tree depth of 10). This discrepancy in burnin time was needed because in preliminary work we observed much quicker convergence with HMC than MCMC sampling. We obtained 1,000 samples from PPDs by saving every second sample for HMC, and 2,000 samples from PPDs for MCMC by saving every fourth sample.

Preliminary inspections of samples showed higher auto-correlation of parameter estimates for MCMC sampling, hence we discarded more samples (higher thinning rate) from the MCMC-derived chains. MCMC convergence was evaluated via the Gelman-Rubin and Geweke statistics (Gelman & Rubin, 1992; Geweke, 1991). We note that the runtime of MCMC could likely be improved by optimizing adaptation, burnin, and sampling steps within JAGS, or by implementing a custom MCMC procedure in the C (or an equivalent) programming language. Data with different dimensions and variance among samples would likely require different optimizations, so we have not further pursued optimization of the MCMC herein. To perform variational inference we used the functionality included within STAN (the 'vb' function; Kucukelbir et al., 2015) and collected 1,000 samples from the estimated posterior distributions.

The ability of models to recover true simulation parameters was estimated via root mean square error (RMSE) and the percentage of times the true simulation parameters were within the 95% high density intervals (HDIs) of PPDs (as per Kruschke, 2015, page 727). For unimodal, symmetric PPDs, the HDI and equal-tailed probability interval should be identical (Gelman et al., 2013, page 38). We measured model bias as the average difference between estimated parameters and the truth and we measured model precision as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, where TP refers to true positives and FP to false positives. False positive rate was calculated as $\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$, where TN is true negatives. Additionally, we calculated Matthew's Correlation Coefficient (MCC; Matthews, 1975), which provides a measure of classifier performance in terms of both true and false positives and negatives. MCC is the correlation between actual and predicted classifications and varies from one (perfect classification) to negative one (completely incorrect classification). An MCC value of zero denotes a classifier that performs no better than expected from random guessing.

2.3 | Data simulation

To evaluate the performance of DMM implementations and alternative statistical methods (see below), we simulated and analyzed data with two sampling categories (k), corresponding to treatment and control groups, or some other blocking factor of interest (Figure 2). We simulated data that possessed three different rank abundance profiles that were meant to correspond to the variety of data encountered by practitioners (Figure 2). We considered simulations in which all features were equally abundant ($1/n$), and two sets of simulations in which features were sampled from Pareto distributions with differing shape parameters. The Pareto distribution describes data with few abundant features and many rarer features (Krishnamoorthy, 2006). The skew towards low abundance in this distribution is controlled by the shape parameter, with smaller parameters increasing skew (Figure 2); the location parameter defines the minimum value of the distribution. For each simulation, we sampled one of these distributions to populate a vector ($\vec{D}$) of length corresponding to the approximate desired number of features (n) within the simulated data:

$$\vec{D} = D_{1\ldots n} = \frac{1}{n}$$
$$\vec{D} \sim \text{ Pareto}\left(\text{shape} = 0.7, \text{location} = 1\right)$$
$$\vec{D} \sim \text{Pareto}\left(\text{shape} = 4, \text{location} = 1\right)$$

$\vec{D}$ was duplicated to make a second vector, $\vec{E}$. Selected features within these vectors were multiplied by an effect size (either 1.1, 1.5, or 2, to simulate 10%, 50%, or 100% shifts in feature relative abundance), such that those elements differed between $\vec{D}$ and $\vec{E}$. Features that varied between vectors were chosen randomly from within each of three broad abundance classes (abundant, rare, and intermediate; see Supporting information) present within $\vec{D}$ and $\vec{E}$. Only features of intermediate abundance were available when constraining all relative abundances to be equal. Effect sizes were applied so that $\sum \vec{D} = \sum \vec{E}$. These two vectors were multiplied by a specified intensity parameter S and used as the parameters for two Dirichlet distributions that were sampled to create $\vec{v}$ parameter vectors for multinomial distributions corresponding with each replicate. In this way, we simulated a replicate by feature matrix where replicates were split into two treatment groups and known features differed between treatment groups. Simulated data sets often had fewer features than the originally specified value for n, because when drawing deviates from multinomial distributions with many rare features, all features would not be observed in each deviate (for a visual depiction of simulation approach see Figure 2).

Using this approach, we simulated data from each sampling distribution that varied in dimensionality (number of features, $\in \left\{500, 2{,}000\right\}$), number of replicates ($\in \left\{10, 50\right\}$), the total number of observations per replicate (e.g., the number of reads per sample for sequencing data; $\in \left\{10{,}000, 50{,}000\right\}$), the variation (noise) among replicates ($\in \left\{0.5, 3\right\}$; the intensity parameter in notation provided above), and the effect size applied to features that differed between sampling groups ($\in \left\{1.1, 1.5, 2\right\}$; to apply the effect size transformation, these values were multiplied by the original proportion). In total, we created and analyzed 144 data sets. Because the same number of observations were used for each replicate, transformation of the data to account for unequal sampling effort was not required. After simulating data matrices, we added a one to every datum, and thereby avoided numerical errors that arise with Dirichlet parameters approaching zero.

For our main simulation, we did not vary read counts among replicates for the sake of simplicity, however to ensure that this did not bias our results we simulated data where replicates differed by up to two orders of magnitude in total observations (read count). To accomplish this, multinomial deviates were obtained as described above, however the total number of draws from the multinomial distribution was randomly selected from $\in \left\{1{,}000, 10{,}000, 100{,}000\right\}$. Data used for this additional analysis were simulated using a representative subset of the aforementioned attributes. Additionally, to better understand the false positive rate of DMM, we simulated and analyzed data where no features were expected to differ between treatment groups, again using a representative subset of the attributes presented above to simulate data.

We competed our implementations of DMM against ALDEX2 v1.14.1 (Fernandes et al., 2014), ANCOM v2.0 (Mandal et al., 2015), DESEQ2 v1.18.1 (Love et al., 2014), EDGER v3.20.9 (Robinson et al., 2010), MVABUND v4.0.1 (Wang et al., 2019) and a frequentist approach using repeated Wilcoxon rank sum tests with a Benjamini-Hochberg false discovery rate (FDR) correction (Weiss et al., 2017). We used multiple comparison correction and typical settings for all software (see Appendix S1). Of the aforementioned methods, only ALDEX2 relies upon DMM. ALDEX2 estimates posterior probability distributions of Dirichlet parameters, which are subsequently transformed via the centered log ratio (Aitchison, 1982). Transformed MCMC samples are subjected to a frequentist test of differential relative abundance between sampling groups, p-values calculated, and the distribution of p-values across MCMC samples obtained (with multiple comparison correction applied as desired by the user). The mean of this distribution is used as a point estimate of the significance of treatment. MVABUND relies on a generalized linear model, in our case using a negative binomial distribution, to determine differential relative abundance. Each feature in the simulated data was a response variable and treatment group was the categorical predictor variable in the model. If the effect of the predictor was significant then the feature differed between treatment groups in relative abundance. MVABUND is thus quite similar to EDGER and DESEQ2, however those methods use different normalization strategies.

Our implementation of DMM differs from these methods in several important ways: (a) most competing methods do not rely on the Dirichlet and multinomial distributions, which explicitly model compositions (except ALDEX2); (b) we use a more complex hierarchical structure than the other methods tested to share information among replicates and sampling groups; (c) we do not perform repeated frequentist tests to determine differences in feature relative abundance, but instead directly subtract posterior probability distributions for parameters of interest and observe the location of zero in the resulting distribution of differences.

For all methods, we evaluated how data attributes (e.g., number of replicates, features, etc.) influenced model performance via multiple regression, with either the proportion of true positives recovered or false positive rate as the response variable.

2.4 | Analyses on empirical data

To understand how DMM could affect inferences made using previously published, empirical data, we analyzed data from Duvallet et al. (2019) describing the lung microbiomes of children with and without oropharyngeal dysphagia (swallowing difficulties) induced aspiration (when a foreign substance enters the lungs). These authors characterized the bacterial assemblages in the lungs (obtained via bronchoalveolar lavage; BAL), gastric fluid, and oropharyngeal region (OR) of each subject via sequencing of the 16S locus. Aspiration is linked to pneumonia in both adults and children (Holas, DePippo, & Reding, 1994; Marik, 2001; Thomson et al., 2016), but the provenance of aspirated microbes is poorly understood. Duvallet et al. (2019) showed that the lung microbiome of patients with difficulty swallowing is more similar to the microbiome of the oropharyngeal region than that of gastric fluid. These authors performed differential relative abundance testing using Kruskal–Wallis tests with a multiple comparison correction to determine whether certain bacterial taxa shifted in relative abundance between aspirating and nonaspirating patients. The authors did not find any taxa that differed in relative abundance, regardless of substrate examined (BAL, gastric fluid, or OR), though they did detect shifts in prevalence (presence across subjects within a sampling group) with phenotype, and suggested that microbial exchange between the lungs and oropharyngeal region is greater than between the lungs and stomach. Using DMM (both VI and HMC; implemented as described above) and all aforementioned competing analyses, we reanalyzed the publicly available BAL data from aspirators and nonaspirators. The data we analyzed were obtained from 66 patients (33 aspirators, and 33 nonaspirators) and included 4,006 OTUs (for details of sequence processing see Duvallet et al., 2019).

3 | RESULTS

Dirichlet-multinomial modelling (DMM) provided a good compromise between true positive recovery and false positive generation (Figure 3; Figure S2), as shown through analysis of data simulated in the context of a treatment-control experimental design. DMM consistently detected many more true positives than competing methods (Figure 4) and this sensitivity facilitated detection of subtle shifts in relative abundance between sampling groups. For instance, when analyzing data with a skewed rank abundance profile, DMM detected ~15%–20% of features that were shifted by treatment by just 10% of their relative abundance. None of the other methods that we employed were able to reliably detect these subtle effects (Figure 3). When effect sizes were larger, DMM recovered more than 80% of true positives on average, which was 20%–40% more true positives than were recovered by DESEQ2, the next best model in terms of sensitivity.

The sensitivity of DMM came at the cost of a slightly higher false positive rate and a loss of precision compared to other methods (Figure 3; Figure S1). Precision was generally high for uniformly distributed data and when the effect size that described the shift in relative abundance of a feature was large, however for data with skewed rank abundance profiles the precision of DMM was lower than competing methods. When considering the Matthew's correlation coefficient (MCC), DMM typically performed as well or better than competing approaches examined (Figure S2). MCC is a more holistic index of classifier performance than precision because it encompasses true and false positives and negatives. MVABUND, ANCOM, and, for some data sets, Wilcoxon tests also performed quite well by this metric.

We observed that the FPR was adversely affected by the rank abundance skew within the data. Analysis of data that was simulated such that no features were expected to differ among treatment groups revealed that for data simulated from a uniform distribution FPR was negligible (0%, Fig. S3). However, FPR for HMC increased to 5.4% on average for data simulated such that they had a highly skewed rank abundance profile (Pareto shape parameter of 0.7). When data were of intermediate skew (Pareto shape of 4) then FPR increased to 8.2%. We also found that high among-replicate variation in sampling depth tended to increase FPR by a few percentage points (Figure S4). On average, FPR of VI was only slightly higher than HMC. By comparison, FPR was often much higher when DMM was implemented via MCMC. Indeed, in many cases, MCMC generated an unacceptably high FPR of over 20%. This high FPR is at least partially due to the lack of convergence we observed for many parameters when using MCMC, even when we employed lengthy run times. We observed broadly comparable results from our primary simulation experiment, which spanned data with a broader variety of attributes and for which features differed in relative abundance among sampling groups (Figure 3).

Of the analytical tools examined, DESEQ2 and EDGER were the next most sensitive behind DMM. DESEQ2 maintained a lower false positive rate than DMM. ANCOM, ALDEX2, and Wilcoxon tests all exhibited negligible false positive rates, but were only able to identify a small fraction of the features that shifted in relative abundance between sampling groups. All methods, including DMM, performed poorly when confronted with data where all features were equally abundant (denoted as 'uniform' in figures). This was unsurprising, because, for these data, the expectation of π was approximately one divided by the number of features present and large, marginal shifts in relative abundance between sampling groups (such as doubling) still resulted in very small differences in proportions (e.g., 1/2,000 vs. 2/2,000), which were difficult to estimate.

We used multiple regression to test how data attributes influenced true positive detection and false positive rate (Tables S2, S3). For all methods competed, the degree of rank abundance skew within the data had, by far, the largest effect on model performance. Surprisingly, all methods were quite insensitive to variation in other data attributes. Data dimensionality (number of features), number of replicates, number of observations, and among-replicate variation

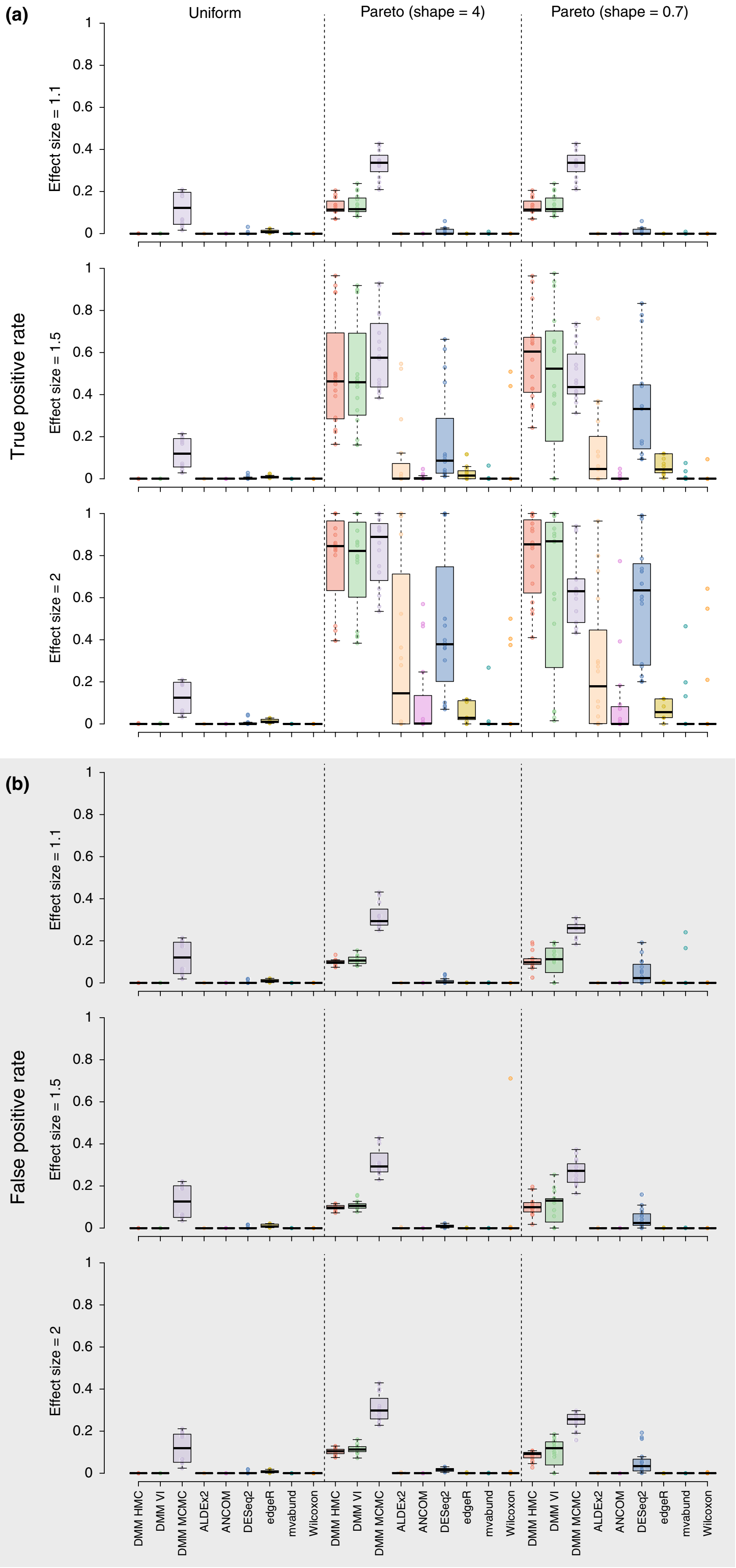


FIGURE 3 Performance of Dirichlet-multinomial modelling (DMM) and competing methods when confronted with simulated data from a treatment-control experimental design. Each point denotes the results from analysis of a simulated data set. Panel a depicts true positive rate and panel b depicts false positive rate. The x axis describes the methods competed, which are each given a unique colour. Each panel is split into three sections that correspond with the three rank abundance profiles used to simulate data (see Figure 2). 'Uniform' refers to data where the expected relative abundance of all features was equivalent; 'Pareto (shape = 4)' refers to data with an intermediate rank abundance skew; 'Pareto (shape = 0.7)' were highly skewed data with very few abundant features and many rare features. Features were made to shift in relative abundance between treatment groups by different effect sizes (an effect size of 1.1 corresponded with a 10% shift in relative abundance). Panels are split by row to show results for a specified effect size. Rectangles in the boxplots delineate the central 50% of the data (first to third quartiles, also called the interquartile range) and contain the median (delineated by a horizontal line). Whiskers extend an additional 1.5 times the interquartile range beyond the first and third quartiles. These are the defaults for boxplots in base R [Colour figure can be viewed at wileyonlinelibrary.com]

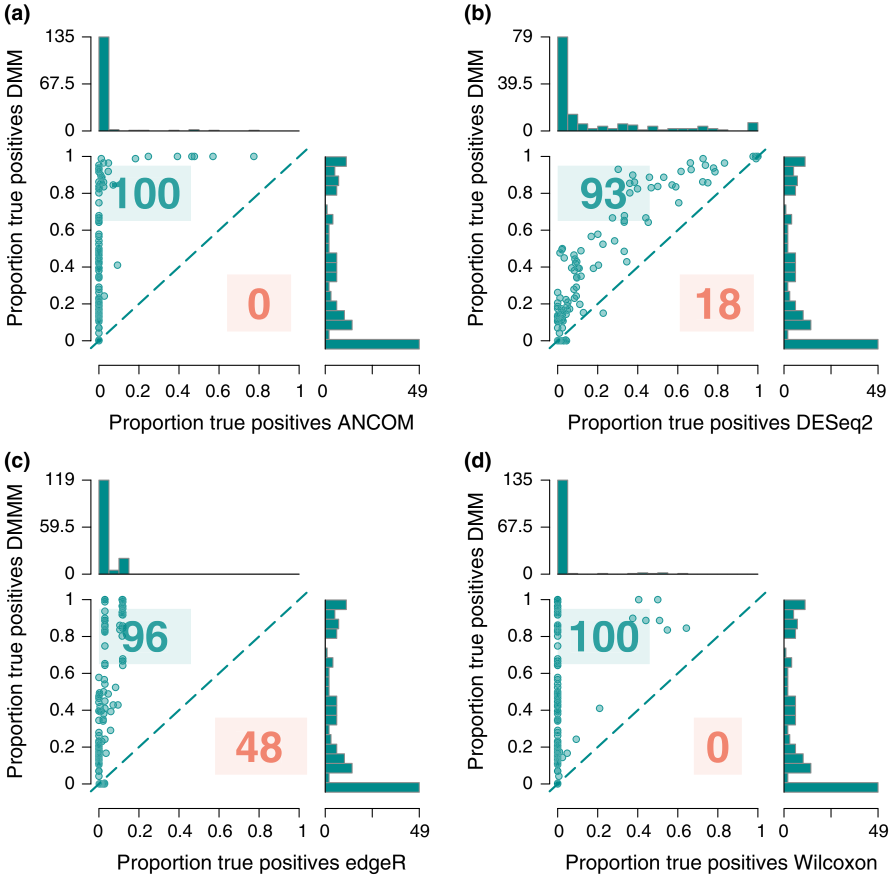


FIGURE 4 Relative ability of competing methods to detect true positives within simulated data. Each point represents the results from a simulated data set and each panel compares the proportion of true positives identified by Bayesian Dirichlet-multinomial modelling (using Hamiltonian Monte Carlo [HMC]) to the proportion of true positives identified by a competing method: (a) ANCOM, (b) DESEQ2, (c) EDGER, (d) Wilcoxon rank sum test. The line bisecting each plot denotes equal performance of both models—so if a point lies above this line then HMC detected more true positives than the competing method for that data set. The summed numbers of points on either side of this line are shown to demonstrate relative performance of methods across datasets. For instance, in panel (a), the Dirichlet-multinomial model (DMM) detected more true positives than ANCOM for 100 data sets, while ANCOM was the more sensitive model for zero data sets. The sum numbers of simulations for each panel differ (and do not always reflect the 144 total data sets analyzed) because in some cases both DMM and the competing method exhibited equal performance. This was mostly the case for extremely challenging data when neither method was able to detect any true positives. Marginal histograms in each plot denote frequency distributions of results along the parallel axis [Colour figure can be viewed at wileyonlinelibrary.com]

had very minor influences on true positive detection and false positive rate for most methods tested (Tables S2, S3).

While our primary goal was ascertaining the relative merits of DMM for detecting differences in feature abundance, we also asked how well DMM could recover the relative abundances ($\vec{D}$ and $\vec{E}$) that were used to simulate data. We report very low average root mean square error (RMSE) for estimates of simulated relative abundances ($\vec{D}$ and $\vec{E}$) obtained through DMM (Figure 5). As a complementary test of model performance, we determined how often the parameters used to simulate data fell within the high density interval (HDI) of PPDs. When feature relative abundances were equal, or modestly skewed ('Equal' or 'Pareto, shape = 4'), the HDI of PPDs encompassed the value used to simulate data for nearly all parameters of interest, regardless of estimation method employed (MCMC, VI, or HMC; Figure S5). Parameter estimation was much more difficult for highly skewed data—when using MCMC or VI, the true values for the parameters did not lie within the estimated HDIs in some cases. By comparison, HMC did better when confronting these challenging data—on average 90% of simulation parameters fell within the HDI, though there was wide variation in model performance depending upon data set (Figure 3). We observed that the width of credible intervals for π parameters was not associated with relative abundance regardless of implementation method or dataset (Figures S16–S18). Bias of DMM differed among implementations, with HMC having negligible bias (Figures S7, S8, S9) and VI and MCMC exhibiting comparatively more bias. We observed that, for all implementations, bias, when present, was typically limited to the most abundant and rarest features within the dataset. Specifically, π parameters were occasionally slightly underestimated for abundant features and overestimated for rare features. This pattern was more noticeable for highly skewed data and can be explained given the prior we used for π parameters, which corresponded to $1/n$, where n was the number of

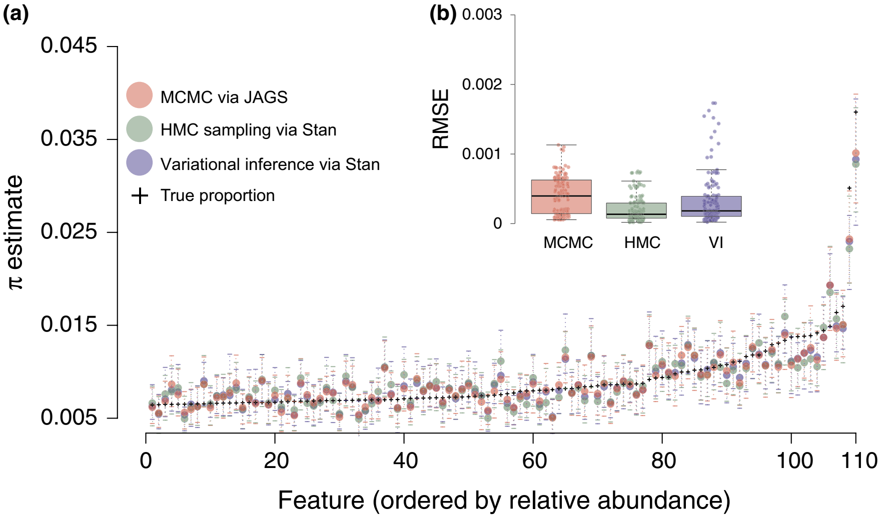


FIGURE 5 Comparison of DMM performance when using different methods to estimate posterior probability distributions (PPDs) of parameters describing feature relative abundance within sampling groups ($\vec{\pi}$; see Figure 1). The results shown in panel (a) are from a single, illustrative simulation. Features are indexed along the horizontal axis and associated $\vec{\pi}$ estimates are shown on the vertical axis. The means of PPDs are shown as shaded circles and the 95% high density interval (HDI) of the PPD is delineated by dotted lines. The true proportions (+ symbols) fall within the HDI of the PPD for almost all features, regardless of PPD estimation method. Average root mean square error (RMSE) for π parameters for all simulated data sets for each method is shown in panel (b) [Colour figure can be viewed at wileyonlinelibrary.com]

features. For skewed data with high among-replicate variation, the strength of the prior was not overcome by the likelihood, thus leading to slight overestimation of marginal features and underestimation of abundant features. If among-replicate variation was reduced, then DMM was able to accurately recover true parameters even for highly skewed data. The prior we chose was agnostic to rank-abundance curves and thus suitable for a wide-range of applications, but could be replaced by a prior with a specific rank-abundance profile if desired by the user.

3.1 | Inferences on empirical data

Reanalysis of data provided by Duvallet et al. (2019) demonstrated the sensitivity of DMM. Using HMC, we found that 53 taxa within the lung microbiome (samples were obtained via bronchoalveolar lavage) shifted in relative abundance between aspirating and non-aspirating children (Figure S19). This contrasts dramatically with the results we obtained from repeated Wilcoxon tests with a Benjamini-Hochberg false discovery rate correction, MVABUND, and ALDEX2, which suggested no taxa significantly shifted in relative abundance between sampling groups. By comparison, DESEQ2 suggested 17 taxa differed, EDGER suggested 10 taxa, and ANCOM four taxa.

Analysis of lung microbiome data using VI and HMC based implementations of DMM provided largely similar results; however, VI did report five fewer taxa shifted in relative abundance than did HMC. The majority of taxa identified by HMC were also identified by VI; the two methods did not agree regarding true positive status for only nine taxa. Of the 53 taxa that we found shifted between sampling groups, the most dramatic change was in a *Streptococcus* taxon, which was much more abundant in aspirating children (Figure S19). An increase in this taxon has previously been reported in adult humans with pneumonia by Akata et al. (2016). We also found an increase in *Haemophilus* (Jacobs & Harris, 1979), *Moraxella* (Claesson & Leinonen, 1994), *Neisseria* (Johnson, Drew, & Roberts, 1981), and *Prevotella* (El-Solh et al., 2003), all of which have previously been associated with pneumonia (see citations for examples), but may be present in healthy lung tissue as well (Beck, Young, & Huffnagle, 2012). We also observed an increase in *Enterobacter*, *Lactococcus*, *Leuoconostoc*, and *Acinetobacter* taxa in the lungs of nonaspirating subjects.

4 | DISCUSSION

Over the past decade, there has been considerable discussion regarding how molecular ecologists should process and analyze compositional data, particularly those generated by high-throughput sequencing instruments (Knight et al., 2018; Thorsen et al., 2016; Weiss et al., 2017). This dialogue has been motivated by the constraints of modern laboratory equipment (e.g., the constant sum constraint of sequencers) coupled with a pressing need for consensus involving appropriate, sensitive tools to analyze data generated by such instruments. Through analysis of simulated data spanning the variation in attributes expected across many scientific domains, we report that new computational statistical techniques have made Dirichlet-multinomial modelling (DMM) an approach that can be applied efficiently in many settings. Specifically, we report that DMM is much more sensitive than the competing approaches we examined, making DMM particularly well suited to identification of subtle shifts in relative abundance among features, such as what might be required in the study of rare, but consequential, microbes or metabolites (Lynch & Neufeld, 2015; Sachdeva et al., 2019). Indeed, for some data, DMM identified many times more true positives than certain competing methods (up to approximately eight times more in extreme cases; Figure 3). The sensitivity of DMM does, however, come at the cost of an increase in false positive rate (FPR) and a loss of precision compared to competing methods, particularly for data with skewed rank abundance profiles and large variation in sampling depth among replicates. For such challenging data, FPR increased to between 5.5%–10% (Fig. S3), which we suggest may be acceptable for those practitioners tasked with analyzing challenging data and who wish to avoid missing features that truly differ among compositions. The tradeoff between sensitivity (also referred to as 'recall') and precision is well known (Buckland & Gey, 1994) and we suggest that the suitability of DMM will depend on the particular needs of the practitioner. If practitioners are interested primarily in sensitivity, then our results suggest DMM is an appropriate method to choose. If, on the other hand, practitioners wish to avoid false positives, even at the expense of considerable loss of sensitivity, then other methods may be more suitable.

Aside from sensitivity, DMM provides several important ancillary benefits including the estimation of parameters that describe the data under consideration and the ability to propagate uncertainty in those estimates to downstream analyses. Propagation of uncertainty allows for a precise statement regarding the credibility of an inference and is a particular benefit of Bayesian techniques over frequentist methods. For example, to determine the extent that specific features shifted from one simulated sampling group to another, we obtained the difference between PPDs of Dirichlet $\vec{\pi}$ parameters from each group (Figure 2d). A PPD is a distribution that explicitly describes the probability of certain values for a particular model parameter; thus, in the model described here, the mean of the PPD for a specific π parameter is a sensible point estimate for that feature's relative abundance and the variation around that mean describes the certainty in that estimate. By subtracting PPDs for π parameters obtained from different sampling groups for a focal taxon, we obtain a PPD of differences, thus propagating uncertainty in relative abundance estimates through to differential relative abundance testing (Figure 2d). This provides a great deal of flexibility to practitioners, because the location of zero in this distribution of differences quantifies the probability that the two original PPDs differed—in other words, that the feature differed in relative abundance between sampling groups. We assumed that, for some feature *i* present in two sampling groups *k*, if 95% of the PPD for $\pi_{ik=1} - \pi_{ik=2}$ does not overlap zero, then that feature differed in relative abundance between groups (see Section 2.1). If a more conservative analysis is desired, then a more strict criterion could be employed to determine if PPDs of focal features are sufficiently divergent, for instance 98% or 99%. Similarly, a less strict criterion could be used (e.g., 90%) for exploratory analyses. Moreover, because we precisely quantify uncertainty in parameter estimates derived from a single model, multiple comparison testing is unneeded for our implementation of DMM. A final benefit of quantifying uncertainty for each feature of interest is that, with some creativity, this uncertainty can be propagated to other downstream analyses, including those using derived parameters of interest such as diversity entropies (see Appendix S1 for how to extract samples from PPDs; Marion, Fordyce, & Fitzpatrick, 2018). The benefits provided by uncertainty propagation are primary differences between DMM as we describe it here and the competing approaches we tested that rely on some form of frequentist testing.

Another important benefit of the approach to DMM we describe is the hierarchical sharing of information among replicates from sampling groups (also see Fordyce et al., 2011). Hierarchical models make thorough use of the information present within the data, which can improve parameter estimates and propagate uncertainty, particularly when sampling effort is inconsistent among replicates and sampling groups (Coblentz et al., 2017). As described in the methods, hierarchical modelling can be used in a way analogous to frequentist, mixed effects modelling to account for nonindependent replicates through the use of a random effect (Bates, Mächler, Bolker, & Walker, 2015; Björk, Hui, O'Hara, & Montoya, 2018). Hierarchical modelling also allows for novel inferential opportunities, given sufficient data, because parameter estimates can be extracted from any level in the model hierarchy.

4.1 | Additional considerations pertaining to Dirichlet-multinomial modelling

A downside to Bayesian modelling is its computational expense. While JAGS (Plummer, 2003), BUGS (Lunn et al., 2012), STAN (Carpenter et al., 2017), and PYMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) have greatly simplified Bayesian model specification and implementation, Bayesian analysis can require much more computation time then frequentist methods. Users should be aware that as the number of parameters to estimate increases, so too does modelling time. For data sets of low to moderate dimensionality (i.e., less than a thousand features), the model described herein can be run on a desktop computer within several hours using any of the three PPD estimation methods (VI may take only a few seconds to run for such small data). However, for larger data sets of many thousand features, convergence when using MCMC or HMC may require a multiple days. For larger data, MCMC sampling should probably be avoided because HMC, as implemented in STAN is much faster and results in convergence for more parameters and, thus, a lower false positive rate (Figure S6). For extremely large data, VI may be the only viable option for efficient parameter estimation. Unfortunately, we observed heightened variation in the performance of VI compared to MCMC or HMC when confronting data with a dramatic rank abundance skew—in some cases VI did as well as HMC, but in other cases it was unable to recover a high proportion of the true positives present (Figure 3). Computational implementations of VI are a topic of current research and will undoubtedly improve over coming years (Blei et al., 2017). For most users, we suggest performing an initial analysis using both HMC and VI. If parameter estimates are largely congruent between techniques (as we generally observed), then VI could be used for subsequent analyses using similar data, thus taking advantage of VI's efficiency.

For HMC or MCMC sampling, time to convergence can be improved through initializing the chains at sensible values for all parameters. We initialize chains for multinomial and Dirichlet parameters at their maximum likelihood values ($\bar{x}_i/N_i$, the proportion of each feature within a sampling group). Additional performance gains can be achieved by combining features that are consistently infrequent across replicates to form a composite feature. This composite feature should be included in modelling, otherwise proportion estimates will be distorted and incorrect. This approach could be particularly appropriate for analysis of high-throughput sequencing of microbiomes and transcriptomes, which often rely on data sets characterized by many features of extremely low relative abundance. Estimates of the relative abundance of very infrequent features will be imprecise, thus precluding effective comparison of relative abundances among sampling groups. Therefore, for some questions, combining these features will not lessen inferential opportunity and can greatly reduce computation time.

Some authors have suggested that the expected negative covariance of feature proportions ($\vec{p}$) in a Dirichlet distribution is a drawback that makes this distribution undesirable (Grantham, Guan, Reich, Borer, & Gross, 2019).; Mandal et al., 2015; Weiss et al., 2016). Specifically, the elements of $\vec{p}$ in a deviate from a Dirichlet distribution are expected to negatively covary (Mosimann, 1963) according to: $Cov\left[p_i,p_j\right] = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$, where $\vec{p}$ is the vector of expected proportions for features in the composition and $\vec{\alpha}$ represents the Dirichlet parameter vector. Indexing of $\vec{p}$ and $\vec{\alpha}$ across features is achieved via i and j, and $\alpha_0 = \sum_{i=1}^{n} \alpha_i$, where n is the number of features. For even modest values of α_0, the expected negative covariance between elements in $\vec{p}$ is small and diminishes rapidly with increasing α_0, approaching zero in the limit of large α_0. The negative covariance structure is a fundamental limitation of compositional data, as one or more features increase, other features must decline to maintain a constant sum. Thus, the Dirichlet distribution assumes a reality that mirrors the data.

There are many problems associated with the analysis of compositional data that cannot be handled by DMM alone (see Aitchison & Egozcue, 2005, Gloor & Reid, 2016, Quinn, Erb, Richardson, & Crowley, 2018, Tsilimigras & Fodor, 2016, van den Boogaart & Tolosana-Delgado, 2013). The most intuitive challenge posed by compositional data is that spurious correlations among features can arise because of the data's inherent covariance structure (Pearson, 1897). For instance, shifts in the relative abundance of a dominant microbial taxon along an abiotic gradient causes shifts in the relative abundance of co-occurring taxa, even if the actual abundances of those taxa are invariant across the gradient (Figure 1). In such a scenario, compositionality could induce associations between the relative abundances of certain taxa and the gradient that are not biologically supported. Other issues that can arise when analyzing compositional data include 'subcompositional incoherence', which means that omission of features from the composition necessarily changes the relative abundances of the remaining features after they are renormalized to their constant sum (e.g., one for proportions; Pawlowsky-Glahn & Egozcue, 2006).

The technique most relied upon to address these problems is log ratio transformation: $\log\left(\frac{p_i}{g(\vec{p})}\right)$, where p_i is the ith feature within $\vec{p}$, which is composed of either counts or proportions, and $g\left(\vec{p}\right)$ is a function. When $g\left(\vec{p}\right)$ is the geometric mean of all feature abundances, this transformation is called the 'centered log ratio' (CLR; Aitchison, 1982). Division by the geometric mean places all replicates on the same scale and, therefore, is useful when variation in sampling effort exists among replicates. Alternatively, $g\left(\vec{p}\right)$ can be an indexing function and output the value of a feature, p_j, that has a constant absolute abundance among replicates. This approach is called the 'additive log ratio' (ALR) transformation (Aitchison, 1982) and can be useful when an internal standard can be added to samples prior to data generation (e.g., during library preparation for next-generation sequencing; Jiang et al., 2011; Munro et al., 2014; Tkacz, Hortala, & Poole, 2018; Tourlousse et al., 2017) or when certain features are expected to be invariant among replicates (e.g., 'housekeeping genes'; Eisenberg & Levanon, 2013). By converting information from each feature into a ratio, both ALR and CLR avoid the subcomposition incoherence problem (Morton et al., 2019). To understand this, consider conducting the ALR transformation on replicates that each include a feature with identical absolute abundance that is used as the denominator in the transformation (it does not matter whether we consider counts or proportions for this example). The ratio between any specific feature within a replicate and the denominator will not be affected by removing other features from the composition (i.e., if the ratio is 2:1 it will remain so after omitting features from the composition and re-normalizing to maintain a constant sum). Either the CLR or ALR transformation can be applied to each MCMC sample of parameters of interest to obtain transformed PPDs for analysis (see Fernandes et al., 2014, for an example).

In conclusion, the challenges posed by many modern molecular ecology data sets—extreme dimensionality, compositionality, and, often, stark differences in the abundance of features—have motivated the rapid development of new analytical tools and techniques. Indeed, new methods and software are published on a near monthly basis and practitioners are left to wonder which tool is best suited for the job at hand. While we do not claim DMM addresses all the challenges associated with compositional data, we do report that it is a sensitive, flexible technique that facilitates feature-specific analyses and should be added to ecologist's toolkits (Fordyce et al., 2011). It is likely to be broadly useful and sensitive for analyses of microbiomes and other datasets generated via DNA barcoding techniques, gene expression, metabolomics, and other applications in molecular ecology (Table S1). To facilitate use of DMM, we have provided an expository vignette in the Appendix S1 that provides an example of how to perform DMM using both STAN and JAGS in the R environment.

The success of DMM for relative abundance estimation, as demonstrated herein, coupled with the aforementioned benefits of hierarchical Bayesian modelling, justifies extension of the DMM to determine the effects of covariates on relative abundances and to characterize mixtures of compositions (sensu Chen & Li, 2013; Holmes et al., 2012; Knights et al., 2011; Shafiei et al., 2015; Tang & Chen, 2018). We look forward to continued method development along these lines.

ACKNOWLEDGMENTS

We wish to thank Claire Duvallet and the coauthors of Duvallet et al. (2019) for making their well-curated data available to the public. Specific thanks to Dr. Duvallet for helpful interpretation regarding our reanalysis of her and her coauthor's data. Additional thanks to James Fordyce and two anonymous reviewers for many helpful comments and criticisms. This research was supported by the Microbial Ecology Collaborative at the University of Wyoming with funding from NSF award #EPS-1655726. Computing was performed in the Teton Computing Environment at the Advanced Research Computing Center, University of Wyoming, Laramie (https://doi.org/10.15786/M2FY47).

AUTHOR CONTRIBUTIONS

All authors contributed to model development and manuscript preparation.

DATA AVAILABILITY STATEMENT

All scripts and processed data used for this manuscript are available at https://github.com/JHarrisonEcoEvo/DMM Harrison, Calder, Shastry, & Buerkle, 2019 and a snapshot corresponding to the status at publication at Zenodo (10.5281/zenodo.3558682). Data from Duvallet et al. (2019) can be downloaded from (https://doi.org/10.5281/zenodo.2678108).

ORCID

Joshua G. Harrison https://orcid.org/0000-0003-2524-0273
W. John Calder http://orcid.org/0000-0002-8923-1803
Vivaswat Shastry https://orcid.org/0000-0002-7294-5607
C. Alex Buerkle https://orcid.org/0000-0003-4222-8858

REFERENCES

Aitchison, J. (1982). *The statistical analysis of compositional data*. New York, NY: Chapman and Hall.

Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, *37*(7), 829–850. https://doi.org/10.1007/s11004-005-7383-7

Akata, K., Yatera, K., Yamasaki, K., Kawanami, T., Naito, K., Noguchi, S., … Mukae, H. (2016). The significance of oral streptococci in patients with pneumonia with risk factors for aspiration: The bacterial floral analysis of 16s ribosomal RNA gene using bronchoalveolar lavage fluid. *BMC Pulmonary Medicine*, *16*(1), 79.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beck, J. M., Young, V. B., & Huffnagle, G. B. (2012). The microbiome of the lung. *Translational Research*, *160*(4), 258–266.

Björk, J. R., Hui, F. K. C., O'Hara, R. B., & Montoya, J. M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, *27*(12), 2714–2724.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science*, *45*(1), 12–19.

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), 94.

Carbonetto, P., & Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, *7*(1), 73–108.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). STAN: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Chen, J., & Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, *7*(1), 418–442.

Claesson, B. A., & Leinonen, M. (1994). *Moraxella catarrhalis*—an uncommon cause of community-acquired pneumonia in Swedish children. *Scandinavian Journal of Infectious Diseases*, *26*(4), 399–402.

Coblentz, K. E., Rosenblatt, A. E., & Novak, M. (2017). The application of Bayesian hierarchical models to quantify individual diet specialization. *Ecology*, *98*(6), 1535–1547.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., … Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683.

Duvallet, C., Larson, K., Snapper, S., Iosim, S., Lee, A., Freer, K., … Rosen, R. (2019). Aerodigestive sampling reveals altered microbial exchange between lung, oropharyngeal, and gastric microbiomes in children with impaired swallow function. *PLoS ONE*, *14*(5), e0216453.

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, *29*(10), 569–574.

El-Solh, A. A., Pietrantoni, C., Bhat, A., Aquilina, A. T., Okada, M., Grover, V., & Gifford, N. (2003). Microbiology of severe aspiration pneumonia in institutionalized elderly. *American Journal of Respiratory and Critical Care Medicine*, *167*(12), 1650–1654.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*, 15.

Fordyce, J. A., Gompert, Z., Forister, M. L., & Nice, C. C. (2011). A hierarchical Bayesian approach to ecological count data: A flexible tool for ecologists. *PLoS ONE*, *6*(11), e26785.

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, *8*(9), e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., … Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Geman, S., & Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In M. A. Fischler, & O. Firschein (Eds.), *Readings in computer vision* (pp. 564–584). San Francisco, CA: Morgan Kaufmann.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Minneapolis, MN: Federal Reserve Bank of Minneapolis, Research Department.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8. https://doi.org/10.3389/fmicb.2017.02224

Gloor, G. B., & Reid, G. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, *62*(8), 692–703.

Grantham, N. S., Guan, Y.,Reich, B. J., Borer, E. T., & Gross, K.. (2019). MIMIX: a Bayesian mixed-effects model for microbiome data from designed experiments. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2019.1626242

Harrison, J. G., Calder, W. J., Shastry, V., & Buerkle, C. A. (2019). Scripts from 'Dirichlet multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data'. https://doi.org/10.5281/zenodo.3558682. Zenodo.

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Holas, M. A., DePippo, K. L., & Reding, M. J. (1994). Aspiration and relative risk of medical complications following stroke. *Archives of Neurology*, *51*(10), 1051–1053.

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE*, *7*(2), e30126.

Jacobs, N. M., & Harris, V. J. (1979). Acute *Haemophilus* pneumonia in childhood. *American Journal of Diseases of Children*, *133*(6), 603–605.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., … Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*. https://doi.org/10.1101/gr.121095.111

Johnson, M. A., Drew, W. L., & Roberts, M. (1981). *Branhamella* (*Neisseria*) *catarrhalis*–a lower respiratory tract pathogen? *Journal of Clinical Microbiology*, *13*(6), 1066–1069. https://doi.org/10.1128/JCM.13.6.1066-1069.1981

Jojic, V., Jojic, N., Meek, C., Geiger, D., Siepel, A., Haussler, D., & Heckerman, D. (2004). Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, *20*(suppl_1), i161–i168.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422.

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., … Kelley, S. T. (2011). Bayesian community-wide cultureindependent microbial source tracking. *Nature Methods*, *8*(9), 761.

Krishnamoorthy, K. (2006). *Handbook of statistical distributions with applications*. Boca Raton, FL: Chapman and Hall/CRC.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R*, JAGS, *and* STAN, 2nd ed. London, UK: Academic Press, Elsevier.

Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in STAN. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *28*, 568–576). Red Hook, NY: Curran Associates Inc.

Logsdon, B. A., Hoffman, G. E., & Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, *11*(1), 58.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESEQ2. *Genome Biology*, *15*, 550.

Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D., Jackson, C., … Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Lynch, M. D. J., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, *13*(4), 217–229.

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

Marik, P. E. (2001). Aspiration pneumonitis and aspiration pneumonia. *New England Journal of Medicine*, *344*(9), 665–671.

Marion, Z. H., Fordyce, J. A., & Fitzpatrick, B. M. (2018). A hierarchical Bayesian model to incorporate uncertainty into methods for diversity partitioning. *Ecology*, *99*(4), 947–956.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451.

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). Methods for normalizing microbiome data: An ecological perspective. *Methods in Ecology and Evolution*, *10*(3), 389–400.

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

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348.

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., … Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*, *10*(1), 2719.

Mosimann, J. E. (1963). On the compound multinomial distribution, the multivariate beta distribution, and correlations among proportions. *Biometrika*, *49*(1/2), 65–82.

Munro, S. A., Lund, S. P., Pine, P. S., Binder, H., Clevert, D.-A., Conesa, A., … Salit, M. (2014). Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nature Communications*, *5*, 5125.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PLoS ONE*, *14*(5), e0216129.

Nowicka, M., & Robinson, M. D. (2016). DRIMSEQ: A Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, *5*, 1356. https://doi.org/10.12688/f1000research.8900.2

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

Pawlowsky-Glahn, V., & Egozcue, J. J. (2006). Compositional data and their analysis: An introduction. *Geological Society, London, Special Publications*, *264*(1), 1–10.

Pearson, K. (1897). Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*(359–367), 489–498.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 10). *124*.

Plummer, M. (2015). RJAGS: Bayesian graphical models using MCMC. R package version 3-15. https://CRAN.R-project.org/package=rjags

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

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EDGER: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rosa, P. S. L., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., … Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, *7*(12), e52078.

Sachdeva, R., Campbell, B. J., & Heidelberg, J. F. (2019). Rare microbes from diverse earth biomes dominate community activity. *bioRxiv*, 636373.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Scordato, E. S. C., Wilkins, M. R., Semenov, G., Rubtsov, A. S., Kane, N. C., & Safran, R. J. (2017). Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. *Molecular Ecology*, *26*(20), 5676–5691.

Shafiei, M., Dunn, K. A., Boon, E., MacDonald, S. M., Walsh, D. A., Gu, H., & Bielawski, J. P. (2015). BioMiCo: A supervised Bayesian model for inference of microbial community structure. *Microbiome*, *3*, 8. https://doi.org/10.1186/s40168-015-0073-x

Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., … Halperin, E. (2019). FEAST: Fast expectation-maximization for microbial source tracking. *Nature Methods*, *1*. https://doi.org/10.1038/s41592-019-0431-x

Stan Development Team. (2018). RSTAN: the R interface to STAN. R package version 2.17.3. http://mc-stan.org

Tang, Z.-Z., & Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, *00*(00), 1–16.

Thomson, J., Hall, M., Ambroggio, L., Stone, B., Srivastava, R., Shah, S. S., & Berry, J. G. (2016). Aspiration and non-aspiration pneumonia in hospitalized children with neurologic impairment. *Pediatrics*, *137*(2), e20151612.

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., … Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, *4*(1), 62.

Tkacz, A., Hortala, M., & Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome*, *6*(1), 110.

Tourlousse, D. M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N., & Sekiguchi, Y. (2017). Synthetic spike-in standards for high-throughput 16s rRNA gene amplicon sequencing. *Nucleic Acids Research*, *45*(4), e23–e23.

Tsilimigras, M. C. B., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Annals of Epidemiology*, *26*(5), 330–335.

van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*. Berlin, Germany: Springer Publishing Company.

Wang, Y., Naumann, U., Eddelbuettel, D., Wilshire, J., Warton, D., Byrnes, J.,… Wright, S. (2019). MVABUND: statistical methods for analysing multivariate abundance data. R package version 4.0.1. https://CRAN.R-project.org/package=mvabund

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., … Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, *10*(7), 1669–1681.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., … Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*, 27.

White, A. E., Dey, K. K., Mohan, D., Stephens, M., & Price, T. D. (2019). Regional influences on community structure across the tropical-temperate divide. *Nature Communications*, *10*(1), 2646.

Yang, Y., Chen, N., & Chen, T. (2017). Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical Bayesian statistical model. *Cell Systems*, *4*(1), 129–137.e5.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Harrison JG, Calder WJ, Shastry V, Buerkle CA. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol Ecol Resour*. 2020;20:481–497. https://doi.org/10.1111/1755-0998.13128