MIDAS: a fast and simple simulator for realistic microbiome data

Mengyu He

mengyu.he@emory.edu

Department of Biostatistics and Bioinformatics,

Emory University, Atlanta, GA 30329, USA

Glen Satten

gsatten@emory.edu

Department of Gynecology and Obstetrics

Department of Biostatistics and Bioinformatics,

Emory University, Atlanta, GA 30329, USA

Ni Zhao*

nzhao10@jhu.edu

Department of Biostatistics,

Johns Hopkins University, Baltimore, MD 21205, USA

Abstract

Motivation: Advances in sequencing technology has led to the discovery of associations between the human microbiota and many diseases, conditions and traits. With the increasing availability of microbiome data, many statistical methods have been developed for studying these associations. The growing number of newly developed methods highlights the need for simple, rapid and reliable methods to simulate realistic microbiome data, which is essential for validating and evaluating the performance of these methods. However, generating realistic microbiome data is challenging due to the complex nature of microbiome data, which feature correlation between taxa, sparsity, overdispersion, and compositionality. Current methods for simulating microbiome data are deficient in their ability to capture these important features of microbiome data, or can require exorbitant computational time.

Results: We develop MIDAS (MIcrobiome DAta Simulator), a fast and simple approach for simulating realistic microbiome data that reproduces the distributional and correlation structure of a template microbiome dataset. We demonstrate improved performance of MIDAS relative to other existing methods using gut and vaginal data. MIDAS has three major advantages. First, MIDAS performs better in reproducing the distributional features of real data compared to other methods at both presence-absence level and relative-abundance level. MIDAS-simulated data are more similar to the template data than competing methods, as quantified using a variety of measures. Second, MIDAS makes no distributional assumption for the relative abundances, and thus can easily accommodate complex distributional features in real data. Third, MIDAS is computationally efficient and can be used to simulate large microbiome datasets.

Availability and implementation: The R package MIDAS is available on GitHub at `https://github.com/mengyu-he/MIDAS`

Contact: Ni Zhao, Department of Biostatistics, Johns Hopkins University (`nzhao10@jhu.edu`)

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The human microbiota and its associated microbiome plays a fundamental role in many diseases and conditions, including obesity (Sze and Schloss, 2016), inflammatory bowel disease (IBD) (Simren et al., 2013), preterm birth (Fettweis et al., 2019), autism (Gilbert et al., 2018) and cancers (Dejea et al., 2014; Kostic et al., 2012). Advances in sequencing technologies, especially 16S rRNA sequencing, now allow rapid and simultaneous measurement of the relative abundance of all taxa in a community. This has lead to a growing number of epidemiological and clinical studies to measure the association between the microbiome and traits of interest, sometimes with complex study designs and research questions.

Although microbiome data is increasingly available, statistical analysis remains challenging. Microbiome data have special characteristics that are difficult to model analytically, including sparsity (the majority of taxa are not present in a sample), overdispersion (the variance of read counts is larger than what is assumed from the usual parametric models), and compositionality (the read counts in a sample sum to a constant). There is little consensus among researchers on how microbiome data should be analyzed, and new methods are being regularly developed, both for identifying individual taxa that associate with diseases (Paulson et al., 2013; Mandal et al., 2015; Lin and Peddada, 2020; Martin et al., 2020; Hu and Satten, 2020; Hu et al., 2021, 2022), and for understanding the community-level characteristics that relate to clinical conditions (Zhao et al., 2015; Wu et al., 2016; Jiang et al., 2022).

Simulating realistic microbiome data is essential for the development of novel methods. To establish the validity of a new method and prove it outperforms existing ones, researchers rely on simulated data in which the true microbiome/trait associations are known. Ideally, the simulated data should be similar to real microbiome data for the simulation studies to be trustworthy. However, simulating realistic microbiome data is made difficult by the same challenges as analyzing microbiome data: sparsity, overdispersion and compositionality. Further, the distribution of counts for each taxon are highly skewed and correlated in a complex way. For these reasons, most simulation methods are based on using a *template* microbiome dataset, and generate simulated data that

is ‘similar’ to the template data in some way.

Several approaches have been proposed for simulating microbiome data. Among them, some methods impose strong parametric assumptions so that the simulated microbiome data share similar dispersion of real data. For example, the Dirichlet-Multinomial (D-M) distribution, in which the taxa counts are generated from a multinomial distribution with proportion parameters provided by a Dirichlet prior (Chen and Li, 2013), is frequently used in simulating microbiome data. The hyper-parameters of this DM model are often estimated from real data so that the simulated data share similar dispersion. Another method, MetaSPARSim (Patuzzi et al., 2019), uses a gamma-multivariate hypergeometric (gamma-MHG) model, in which the gamma distribution models the biological variability of taxa counts, accounting for overdispersion, and the MHG distribution models technical variability originating from the sequencing process. Although the D-M model and the MetaSPARSim model address the compositional feature by either the multinomial or the hypergeometric distribution, they do not attempt to match the correlation structures in the simulated data with those found in the real data.

One recently-developed approach that does attempt to model between-taxa correlations is SparseDOSSA (Sparse Data Observations for the Simulation of Synthetic Abundances) (Ma et al., 2021). This hierarchical model makes assumptions about both the marginal and joint distributions of the relative abundances of a set of taxa. For the marginal distribution, SparseDOSSA assumes a zero-inflated log-normal model for the relative abundance of each taxon and then imposes the compositional constraint. Parameters in the zero-inflated log-normal marginal are estimated through a penalized Expectation-Maximization (EM) algorithm from a template dataset. Unfortunately, the penalized EM algorithm for estimating hyper-parameters is computationally expensive, especially when a large number of taxa exist in the data. For example, fitting SparseDOSSA model to a modest-sized dataset with sample size of 79 and number of taxa = 109 takes more than a day ($\approx$ 27.8 hours) on a single Intel “Cascade Lake” core (Ma et al., 2021). To partially compensate for this drawback, SparseDOSSA provides fitted models that were previously trained by the developers and that users can use directly, which is only useful if the developer-provided fits resemble

the data users wish to generate. Moreover, SparseDOSSA removes rare taxa that appear in fewer than 4 samples by default, thus failing to accommodate the possibility that rare taxa are of interest in the simulation studies.

Considering the drawbacks of existing approaches, a method that can flexibly capture the distributional and correlation structure of microbiome data would greatly benefit the research community. In this paper, we develop a fast and simple MIcrobiome DAta Simulator (MIDAS) for generating realistic microbiome data that capture the correlation structure of taxa of a template microbiome dataset in both the presence-absence data and the relative abundances. MIDAS generates relative abundance data using a two-step approach. The first step generates the presence-absence of each taxon by simulating correlated binary data from a probit model with a shrinkage correlation structure, while the second step generates relative abundance and count data from a Gaussian copula model. MIDAS also allows the user to change the library sizes, taxon relative abundances or the proportion of non-zero cells to generate data in which these features may depend on covariates such as case/control status.

MIDAS uses the correlation structure of the template data both in generating correlated binary data for the first step, and in the Gaussian copula model for relative abundances. A rank-based approach is used to fit the copula model to the template data to handle zero count data. MIDAS makes no distributional assumption for the relative abundances, and thus can accommodate complex distributional features in real data. Using simulations, we show that MIDAS reproduces the distributional features at both the presence-absence level and the relative-abundance level, and in so doing generates data that are more similar in multiple metrics to the template data than competing methods. MIDAS is also computationally efficient and can be used to simulate large microbiome datasets in a fast and simple fashion.

2 Materials and methods

Suppose that we are interested in simulating microbiome data that are similar to a real dataset with n samples and J taxa; we refer to this dataset as the 'template' data. Assume in this dataset, all taxa are present in at least one sample, and J_1 taxa ($J_1 \leq J$) are present in all samples. For sample i and taxon j, let C_{ij} denote the observed count, $N_i = \sum_{j=1}^{J} C_{ij}$ denote the observed library size, π_{ij} denote the observed relative abundance ($\pi_{ij} = C_{ij}/N_i$), and let presence-absence indicator $Z_{ij} = \mathbb{I}(C_{ij} > 0)$ where $\mathbb{I}(S) = 1$ if S is true and 0 otherwise. We let $\mathbf{C}$, $\mathbf{Z}$ and $\boldsymbol{\pi}$ represent the $n \times J$ matrices of the read counts, presence-absence and the relative abundances of all taxa in the template data, respectively. Corresponding quantities for the simulated data are denoted by a tilde, e.g. $\widetilde{\mathbf{Z}}$ is the presence-absence indicator in the simulated data.

We develop a two-step procedure for generating count and relative abundance data that share similar characteristics of a template dataset. The first step is to generate the binary presence-absence indicators so that they share similar correlation as in the real presence-absence data $\boldsymbol{Z}$. This step allocates zeros and non-zeros for the simulated data. The second step is to fill the non-zero cells from step 1 using a Gaussian copula model fitted to the observed values $\boldsymbol{\pi}$. We next describe each step in detail; summaries of each step are found in algorithm 1.

2.1 Step 1: generate presence-absence data

The goal of step 1 is to generate presence-absence data $\widetilde{Z}_{ij}$ having correlation and marginal means that match the target data. To facilitate the generation of these data, we propose to generate multivariate normal data D_{ij} having mean $\mu_j + \eta_i$ and variance-covariance matrix Σ in such a way that $Z_{ij} = 1$ corresponds to $D_{ij} \geq 0$. To accomplish this, we choose μ_j and η_i to jointly solve

$$\sum_{j=1}^{J} \Phi(\mu_j + \eta_i) = Z_{i\cdot} \tag{1}$$

and

$$\sum_{i=1}^{N} \Phi(\mu_j + \eta_i) = Z_{\cdot j} \ , \tag{2}$$

where $Z_{i\cdot}$ is the number of non-zero cells in the data from the ith observation and $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the CDF and quantile function of the standard normal distribution respectively. These equations are iterated alternately, starting from the initial values $\eta_i = 0$ and $\mu_j = \Phi^{-1}(Z_{\cdot j})$.

To estimate Σ we first calculate the empirical correlation matrix for the observed values of $\mathbf{Z}$, denoted ζ. We next convert ζ into a matrix of tetrachoric correlations, denoted by ρ, using the approach of (Bonnet and Price, 2005). The correlation matrix is smoothed to be positive definite using the function `cor.smooth()` in R package `psych` (Revelle, 2015). We then sample values $\widetilde{D}_{ij} \sim \mathrm{MVN}(\mu_j + \eta_i, \rho)$ and take $\widetilde{Z}_{ij} = \mathbb{I}(\tilde{D}_{ij} > 0)$.

Because the number of zero cells in a sample is related to its library size, this procedure is designed so that the library sizes in the simulated data are the same as the template data. See Section 2.3 for discussion of the case where simulated data with different library sizes than those in the template data is required.

2.2 Step 2: generate relative abundance and count data

The goal of Step 2 is to generate relative abundance data that mimic the relative abundance data of the non-zero cells in the target data. To accomplish this, we use a Gaussian copula model, which allows us to specify a marginal distribution for each taxon that matches the observed distribution of non-zero relative abundances for that taxon.

In order to allow for the possible generation of non-zero relative abundances for taxa that are observed to have zero counts, we must include the zero cells when we specify the correlation structure of the Gaussian copula. To accomplish this, we use a rank-based approach based on the relationship between the Pearson and Spearman correlations for normally-distributed data (Ruppert and Mattesson, 2015). This approach does not require us to know the values we would have obtained for an empty cell, had that cell not been empty; our only assumption is that the relative

abundances of the zero cells are smaller than those of the cells having non-zero counts. In particular, to specify the correlation of the underlying Gaussian model, we calculate Spearman's rank correlation ϕ for the observed relative abundance values. When calculating the rank correlation, we consider the zero cells to be tied, and then break these (and any other) ties by a random ordering. For the kth of K such random orderings, after computing Spearman's rank correlation $\phi^{(k)}$, we obtain the corresponding Pearson correlation $r^{(k)}$ using $r_{ij}^{(k)} = 2sin(\pi\phi_{jj'}^{(k)}/6)$. The correlation matrix $r^* = \sum_{k=1}^{K} r^{(k)}/K$ is corrected to be positive definite by setting negative eigenvalues to a small positive value and then renormalizing to preserve the trace of the smoothed correlation matrix. The default choice for MIDAS is $K = 100$. We then take the corrected correlation matrix as the final correlation matrix for the underlying Gaussian model.

To simulate a new dataset with n observations, we first generate n independent multivariate normal variables $W \sim \mathrm{MVN}(0, r^*)$ from the multivariate normal distribution with mean $\mu = 0$ and variance-covariance matrix r^*. We then choose simulated relative abundances for the jth taxon, $\widetilde{\pi}_{\cdot j}$, from the non-zero relative abundances in the template data in the following way. If $\widetilde{Z}_{ij} = 0$ we choose $\widetilde{\pi}_{ij} = 0$. For the remaining $\widetilde{m}_j = \sum_{i=1}^{n} \widetilde{Z}_{ij}$ values, if $\widetilde{m}_j \leq m_j$ where $m_j = \sum_{i=1}^{n} Z_{ij}$, we sample $\widetilde{m}_j$ values from the non-zero values of $\pi_{\cdot j}$ without replacement; if $\widetilde{m}_j > m_j$ then, in addition to the m_j non-zero values of $\pi_{\cdot j}$, we sample an additional set of $\widetilde{m}_j - m_j$ values from the non-zero values of $\pi_{\cdot j}$ with replacement. We then assign the non-zero values of $\widetilde{\pi}_{ij}$ the values of π_{ij} that we have sampled, in such a way that their rank agrees with the rank of the $w_{\cdot j}$ values corresponding to $\widetilde{Z}_{ij} = 1$. The values of $\widetilde{\pi}$ are then normalized to sum to one for each observation.

A count table $\widetilde{C}$ is then calculated by multiplying the sampled relative abundances $\widetilde{\pi}_{ij}$ by library size N_i for each observation. Any values so obtained that are between 0 and 1 are rounded up to 1 to keep the presence-absence structure; other values are rounded to the nearest integer. The library sizes for the simulated data are then calculated as $\widetilde{N}_i = \sum_{j=1}^{J} \widetilde{C}_{ij}$ and the final relative abundance is updated through $\widetilde{\pi}_{ij} = \widetilde{C}_{ij}/\widetilde{N}_i$.

Algorithm 1 Steps to simulate one data set similar to a template data

Input: Template data: presence-absence matrix $\mathbf{Z}$ $(n \times J)$, relative abundance matrix $\boldsymbol{\pi}$ $(n \times J)$, K (number of Monte-Carlo replicates for $\mathbf{r}^*$)

Output: $\widetilde{\mathbf{Z}}$,$\widetilde{\mathbf{C}}$, $\widetilde{N}_i = \sum_{j=1}^{J} \widetilde{C}_{ij}$, $\widetilde{\pi}_{ij} = \mathbf{C}_{ij}/N_i$, and $\widetilde{m}_j = \sum_{i=1}^{n} \widetilde{Z}_{ij}$

Initialize necessary quantities for presence-absence data

a. Filter $\mathbf{Z}$ to remove taxa that are present in all samples.

b. Calculate the correlation matrix ζ using the filtered $\mathbf{Z}$.

c. Convert ζ to the tetrachoric correlation matrix (with shrinkage to remove negative eigenvalues) to obtain ρ.

d. Estimate μ and η

Initialize necessary quantities for relative abundance data

a. **for** $k = 1, \cdots, K$,

Calculate $\phi^{(k)}$ = Spearman's rank correlation of $\boldsymbol{\pi}$ by randomly breaking ties.

Convert $\phi^{(k)}$ to Pearson correlation $\mathbf{r}^{(k)}$

Calculate $\mathbf{r}^* = \sum_{k=1}^{K} \mathbf{r}^{(k)}/K$.

Step 1: Generate presence-absence data $\widetilde{\mathbf{Z}}$.

a. Set $\widetilde{\mathbf{Z}} = 1$ for taxa present in every observation.

b. For remaining taxa, generate multivariate normal data $\boldsymbol{D}_{ij} \sim \mathrm{MVN}(\mu_i + \eta_j, \rho)$.

c. Set $\widetilde{\mathbf{Z}}_{ij} = I(D_{ij} > 0)$ for taxa not present in all observations.

Step 2: Generate relative-abundance and count data.

a. Generate $\mathbf{W} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{r}^*)$.

b. **for** $j = 1, \cdots, J$,

Set $\widetilde{\pi}_{ij} = 0$ for the $n - \widetilde{m}_j$ observations in taxon j having $\widetilde{Z}_{ij} = 0$

Sample $\widetilde{m}_j$ elements from non-zero elements from $\pi_{\cdot j}$

Assign sampled values to non-zero $\widetilde{\pi}_{ij}$ in order given by W_{ij}.

c. Normalize the matrix with assigned relative abundances.

d. Estimate counts by multiplying matrix of assigned relative abundances by N_i for each observation; Round non-zero counts less than 1 up to 1; round all other counts to nearest integer to yield count matrix $\widetilde{\mathbf{C}}$.

e. Calculate library size $\widetilde{N}_i = \sum_{j=1}^{J} \widetilde{\mathbf{C}}_{ij}$.

g. Calculate final relative abundance matrix $\widetilde{\boldsymbol{\pi}}_{i\cdot} = \widetilde{\mathbf{C}}_{i\cdot}/\widetilde{N}_i$.

2.3 Changing the parameters of the simulation

In the previous sections, we have shown how MIDAS will generate a dataset that replicates a template microbiome dataset. In some situations, however, we may want to change the parameters used in MIDAS to generate simulated data that are different from the template data in a controlled way. For example, we may want to generate data with different library sizes. Or, we may want to change the relative abundances or proportion of zero cells for a set of taxa, perhaps in a way that is related to external covariates. In this section, we discuss how this can be accomplished in the MIDAS framework.

The parameters that are available for adjustment in MIDAS are: the library sizes N_i, the taxon mean relative abundances p_j, the taxon proportion of zero cells $\delta_j = Z_{\cdot j}/n$ and the taxon mean relative abundance among non-zero cells $p_j^{(1)}$. Note that these last three are related through

$$p_j = \delta_j \times p_j^{(1)}. \tag{3}$$

We denote the new (desired) parameters as $\widehat{N}_i$, $\widehat{p}_j$, $\widehat{\delta}_j$ and $\widehat{p}_j^{(1)}$. Aside from the two correlation matrices (which do not change), the quantities that govern the MIDAS simulation are $Z_{i\cdot}$, $Z_{\cdot j}$ and the values of π_{ij} for non-zero cells. Thus, we must describe how these quantities are obtained when a change in parameters is desired. MIDAS supports two approaches to changing parameters. In the first approach, the user specifies changes to at most one taxon-level quantity and optionally the library sizes, and MIDAS changes the other parameters to be consistent with the patterns seen in the target data. In the second strategy, the user specifies all the quantities. The first should generate more realistic data, but the second will be useful for generating data where only a single quantity has been changed, which is useful when evaluating the performance of microbiome analysis methods. We first describe implementation of the first strategy which is used for the rest of this work; the second strategy is described briefly at the end of this section.

For realistic data when changing library sizes, it is important to account for the positive relationship between the library size and the number of taxa with non-zero counts for that sample (see

e.g. Supplementary Figure S1). To accomplish this, MIDAS fits the Shape Constrained Additive model (SCAM) (Pya and Wood, 2015)

$$log_{10}[\mathbb{E}(Z_{i\cdot})] = f(log_{10}(N_i)), \tag{4}$$

to the template data using the R package `scam`, where f is a monotone smoothing spline. Then, if we wish to generate data having library sizes $\widehat{N}_i$ we can generate $\widehat{Z}_{i\cdot}$ using (4). We would then replace the right-hand side of (1) by $min(J, \widehat{Z}_{i\cdot})$ and replace the right-hand side of (2) by $\widetilde{Z}_{\cdot j} = (Z_{\cdot j}/Z_{\cdot\cdot})\widehat{Z}_{\cdot\cdot}$ where $\widehat{Z}_{\cdot\cdot} = \sum_i min(J, \widehat{Z}_{i\cdot})$ to find the appropriate values of μ and η to simulate presence-absence data $\widetilde{Z}$.

Many simulations require groups of samples with different relative abundances. When changing the mean relative abundance of the taxa, we can expect the proportions of zero cells to change as well. To account for this, MIDAS fits the SCAM model

$$logit\left[\mathbb{E}\left(\delta_j\right)\right] = g(log_{10}(p_j)) \tag{5}$$

to the template data. Then realistic values of $\widehat{\delta}_j$ and hence $Z_{\cdot j}$ can be generated for the specified $\widehat{p}_j$ values. Alternatively, a simulation may specify a new value of $\widehat{\delta}_j$. In this case MIDAS determines the resulting mean relative abundances using

$$logit\left[\mathbb{E}\left(p_j\right)\right] = h(log_{10}(\delta_j)) \tag{6}$$

In either case, the constraint (3) then determines $\widehat{p}_j^{(1)}$. MIDAS then replaces the right-hand side of (2) with $\widehat{Z}_{\cdot j}$ and changes the right-hand size of (1) to $(Z_{i\cdot}/Z_{\cdot\cdot})\widehat{Z}_{\cdot\cdot}$ where $\widehat{Z}_{\cdot\cdot} = \sum_j \widehat{Z}_{\cdot j}$ to find the appropriate values of μ and η.

Finally, if we wish to change *both* library sizes and taxon frequencies, we fit the SCAM model

$$logit[\mathbb{E}(Z_{ij})] = f(log_{10}(N_i)) + g(log_{10}(p_j)) \tag{7}$$

then use the marginal totals $\widehat{Z}_{i\cdot}$ and $\widehat{Z}_{\cdot j}$ predicted from (7) in (1) and (2). Simultaneous changes in both library sizes and the proportion of non-zero cells are not currently supported by MIDAS.

To allow the user more control over the simulation when parameter changes are made, MIDAS supports a second simulation strategy in which the user specifies N_i and (at least) two of $\widehat{p}_j$, $\widehat{\delta}_j$ and $\widehat{p}_j^{(1)}$. Then, the SCAM models described above are not fit; instead, $\widehat{Z}_{\cdot j} = n\widehat{\delta}_j$ after which MIDAS assigns $\widehat{Z}_{i\cdot} = (Z_{i\cdot}/Z_{\cdot\cdot})\widehat{Z}_{\cdot\cdot}$ where $\widehat{Z}_{\cdot\cdot} = \sum_j \widehat{Z}_{\cdot j}$.

Changes to $Z_{i\cdot}$ and $Z_{\cdot j}$ affect only generation of the binary data $\widetilde{\mathbf{Z}}$. To ensure generation of relative abundance data that agrees with $\widetilde{p}_j^{(1)}$ for either strategy, MIDAS replaces by $\pi_{ij}^{\alpha_j}$ where α_j is chosen so that their mean among non-zero cells is $\widehat{p}_j^{(1)}$. Presence-absence data $\widetilde{\mathbf{Z}}$ and relative abundance data $\widetilde{\boldsymbol{\pi}}$ are then generated as described in Sections 2.1 and 2.2, and are scaled by $\widehat{N}_i$ to generate counts $\widetilde{\mathbf{C}}$ as described in Section 2.2.

Finally, it may be worth noting that we may choose to use only a subset of the template data. For example, if we wish to compare 'control' and 'case' populations, we could use only disease-free observations from the template data, so that the proposed changes in taxon prevalences represent modifications to the disease-free group to create 'case' data.

2.4 Simulation studies and comparisons with existing methods

We compared MIDAS to three competing methods (the D-M method, MetaSPARSim and SparseDOSSA) and evaluate how well the simulated data reporduce the characteristics of the template data. We use two datasets from the Integrative Human Microbiome Project (HMP2) (Proctor et al., 2019) as the template data: a vaginal microbiome dataset from Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI) project, and a gut microbiome dataset from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB) project (Lloyd-Price et al., 2019). These two datasets represent microbiome from two body sites that are frequently studied in the literature. They are also distinct in their characteristics, and thus provide a comprehensive assessment of the proposed method. For example, the vaginal data, containing 95.25% zeros, is more sparse than the gut data, which comprises 85.09% zeros. Moreover, the coefficient of variation (CV) of

vaginal data is 40.77, while that of the gut data is 10.76, indicating that the vaginal data is more over-dispersed.

We lightly filtered the two template datasets. For quality control, we removed samples with library size < 3000. To allow comparison with SparseDOSSA, we removed taxa that were present in fewer than 4 samples, a requirement of SparseDOSSA. The MOMSPI was a longitudinal study with repeated vaginal samples; we kept only first-visit samples to avoid repeated measures. The only filtering used for the IBD data was that required by SparseDOSSA. After filtering, 517 samples and 1146 taxa were preserved in the vaginal MOMS-PI dataset; the gut IBD dataset comprised 146 samples and 614 taxa. We ignored covariates such as gender or location of biopsy collection to focus only on reproducing the microbiome datasets as closely as possible, the goal of all methods considered here. In our simulations, the library sizes for datasets generated using the D-M method and MetaSPARSim were the same as that in the original data. For SparseDOSSA, the library sizes were generated from a log-normal distribution parameterized by mean and standard deviation of log counts in the original data, as recommended in their original publication. To facilitate comparison of the methods, all simulated counts were transformed to relative abundances.

We compared the simulated data from each method to the template data using several measures. First, we concatenated the template data with a simulated dataset from each method, and defined a binary variable to differentiate the template and simulated data. We tested the significance of this variable using PERMANOVA (Anderson, 2001), which tests for shifts in the between-observation distances. Our PERMANOVA tests used the Jaccard distance as well as the Bray-Curtis distance, which are both commonly used in microbiome data analyses. The Jaccard distance uses only presence-absence information in the data, and thus can assess how similar $\widetilde{\mathbf{Z}}$ and $\mathbf{Z}$ are, while the Bray-Curtis distance accounts for both the presence-absence and relative abundance information and can be used to assess the simulation of $\widetilde{\pi}$. We also compared the alpha diversity of the simulated data and template data. The simulated communities were compared to the template in terms of observed richness and Shannon Index, and the differences in diversity were tested by Kruskal-Wallis tests. The observed richness is simply the number of observed taxa, while Shannon

Index additionally considers evenness-the relative abundances of taxa-when quantifying diversity. To suppress random variability, we repeated the comparison of alpha-diversity and beta-diversity using 20 simulated datasets from each of the four methods. Finally, we compared the methods visually, using ordination and PCoA, as well as boxplots of alpha diversity values, using a single simulated data set for each method.

We next compared the simulation approaches in terms of their β-dispersion, by comparing whether the distribution of distances from each observation to the sample centroid was the same in the simulated and template data. We calculated distances to the centroids using the betadisper function in R package `vegan` (Dixon, 2003). We used the Kolmogorov-Smirnov (K-S) test to compare these empirical distributions. We again averaged results over 20 simulation replicates to suppress random variability. We also compared the alpha diversity of the template and simulated data, as measured by the species richness (number of observed taxa) and the Shannon entropy.

Finally, we evaluated the performance of our approach to generating data with different library sizes by rarefying our template datasets, then using the approach described in section 2.3 to increase the library size to that of the original template data. Thus, we can compare the resulting simulated data to the original template data. Specifically, for each template, the observed counts for each subject were rarefied (subsampled without replacement) to remove 10% of the observed counts. The rarefied data are then treated as the template data in MIDAS, and the target library size is the original library size.

3 Results

3.1 Comparison Results

The PCoA plots in Figure 1 provide a simple visualization of the similarities between the original data and the simulated data by MIDAS, the D-M method, MetaSPARSim, and SparseDOSSA for the IBD data and MOMS-PI data. For both datasets, after ordination the data simulated from MIDAS looked similar to the template data, using either the (presence-absence-based) Jaccard

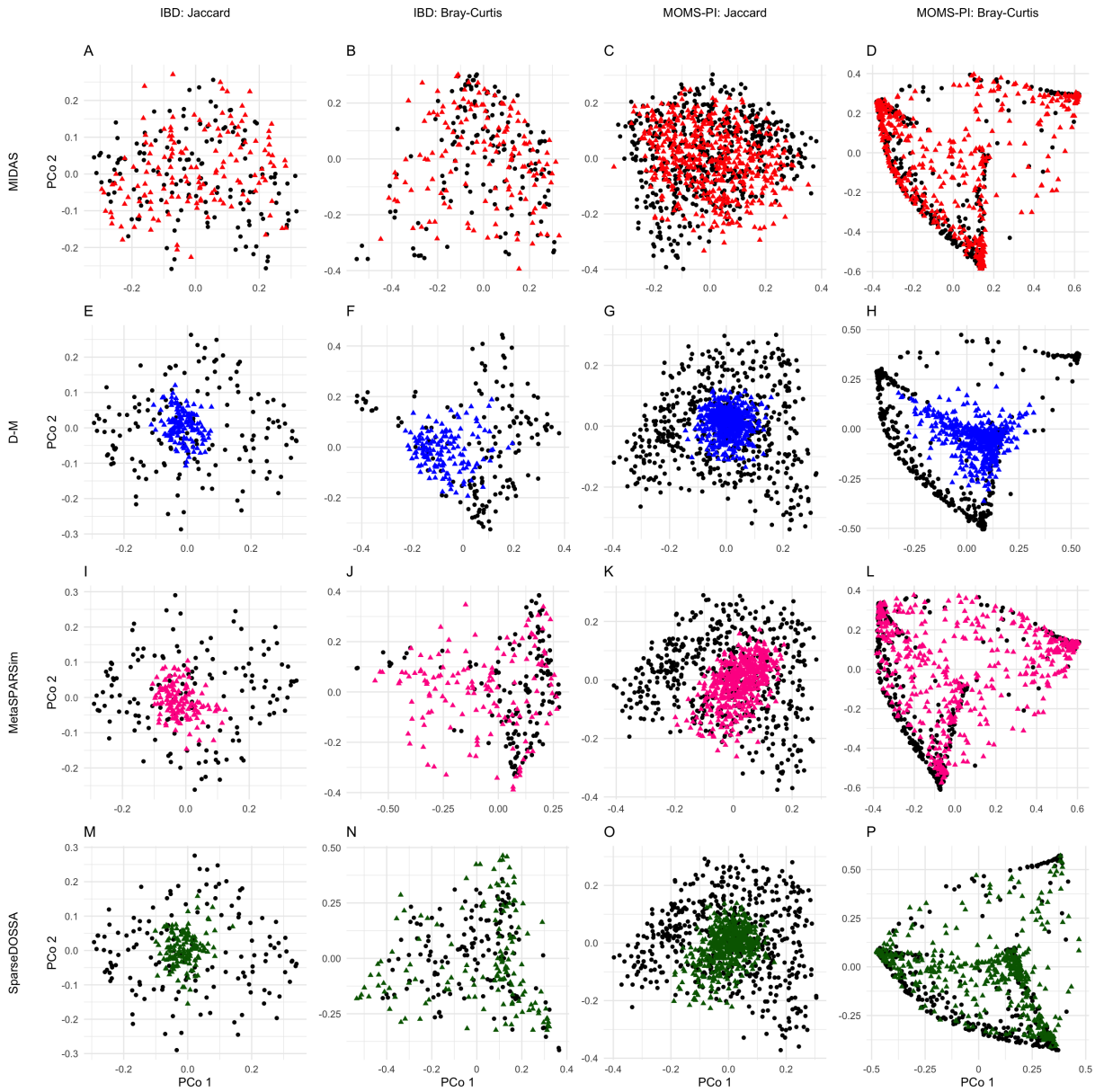

Figure 1: Principal Coordinates plots (PCoA) of the simulated and original community. Each row corresponds to one method. The left two columns are the plots for the IBD data, and the right two columns are the plots for the MOMS-PI data. Black points: samples from original data. Colored points: samples from the simulated data with red being MIDAS, blue being D-M, pink being MetaSPARSim, and green being SparseDOSSA

(Figure 1 A,C) or (relative abundance-based) Bray-Curtis distance (Figure 1 B,D). To allow visual comparison between the template data and multiple datasets simulated by MIDAS, in Figure 2 we also give a heatmap of constructed using 20 simulated datasets. Conversely, for both data

templates, data simulated by the D-M method, MetaSPARSim, SparseDOSSA all appear to be underdispersable in the first two principle coordinates (Figure 1 E,G,I,K,M,O) using the Jaccard distance. For the IBD data, data simulated using D-M and MetaSPARSim appeared easily distinguished from the original data when the Bray-Curtis distance was used (Figure 1 F, J). For the MOMS-PI data, we also see clear underdispersion in data simulated using D-M (Figure 1 H).

The visual impressions of beta diversity in figures Figure 1 and Figure 2 are confirmed in Table 1, where we test whether the template and simulated data are significantly different using PERMANOVA. For tests using the Jaccard distance, the p-values for MIDAS were consistently high (indicating no detected difference between simulated and template data); for the other methods, only SparseDOSSA showed a similar pattern to the template data, but only when applied to the IBD data. When using the Bray-Curtis distance, only MIDAS could produce data that was not easily differentiated from the template data by PERMANOVA.

Figure 3 shows empirical cumulative distribution function (CDF) representing the distances between each sample and the group centroid in the simulated data and in the template data using Jaccard and Bray-Curtis distances, calculated using the `betadisper` function in the R package `vegan`. If the simulated data are similar to the template data, the CDF should resemble that of the template data. The CDFs of the template data and the datasets simulated by D-M method, MetaSPARSim, and SparseDOSSA are noticeably dissimilar, which is confirmed by extremely small Kolmogorov-Smirnov test p-values. The range of distances to centroids in the simulated data by D-M method and SparseDOSSA is smaller compared to the real data in every scenario, indicating a smaller dispersion overall. For the IBD data, the MIDAS-simulated data follows the template data closely in dispersion in both Jaccard and Bray-Curtis distances. For the MOMS-PI data with Bray-Curtis distance, there is a significant difference in dispersion between the MIDAS-simulated data and the template data Figure 3. However, panel D of Figure 3 shows the MIDAS results are clearly closer to those of the template data than the other methods are.

Table 1 and Figure 4 also shows comparisons of two measures of alpha diversity: species richness and Shannon index. To compare the richness and Shannon index between simulated and

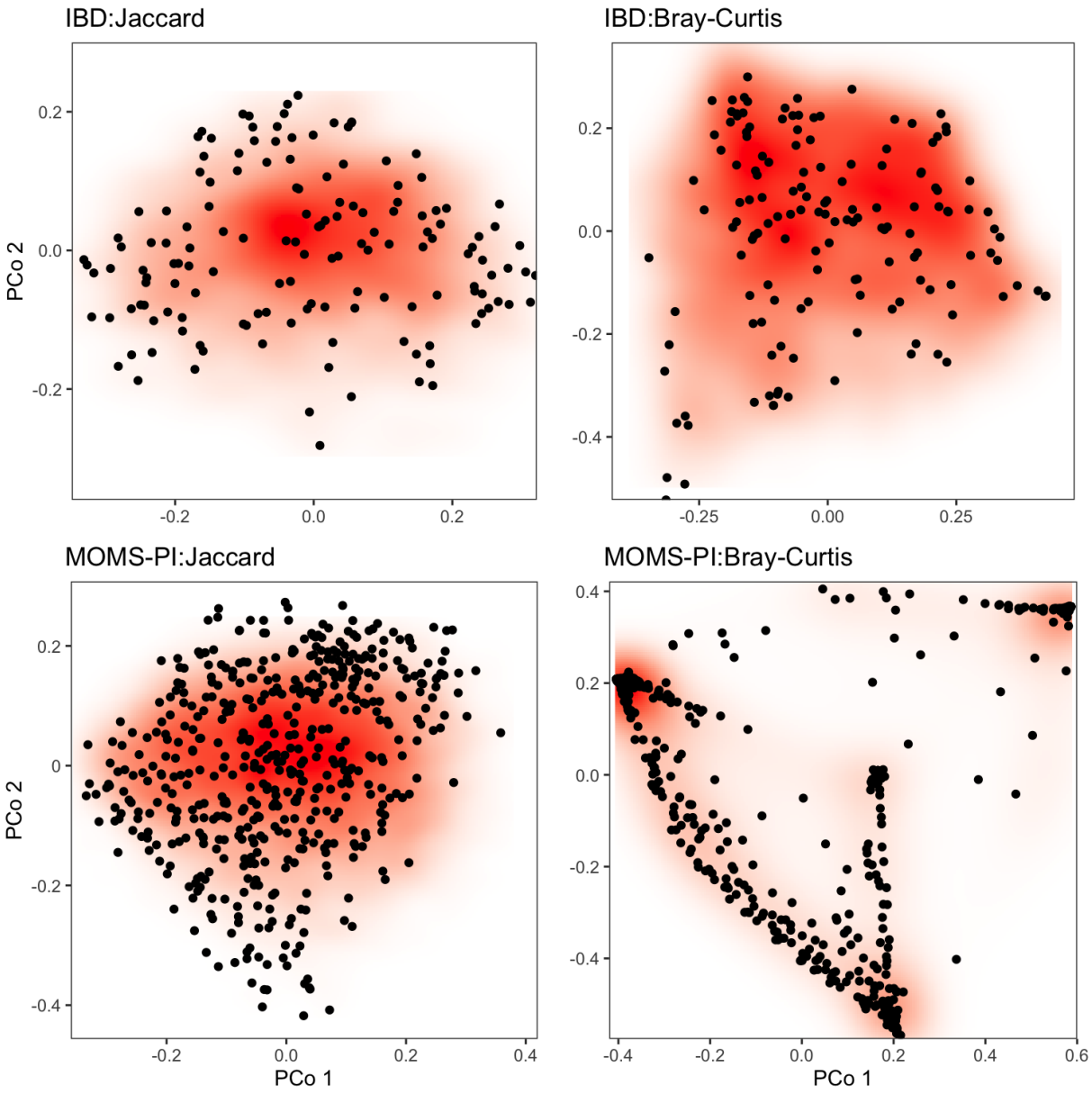

Figure 2: Principal Coordinates plots (PCoA) of simulated and original community. The heatmap is plotted based on 10 replicates of simulated communities by MIDAS, with darker colorings associated with higher density of simulated values. Black points represent the original community.

Table 1: Average p-value (from 20 replicates) for tests comparing alpha and beta diversities of simulated data and template data

		Beta-Diversity		**Alpha-Diversity**			
Data	**Method**	**Jaccard**	**Bray-Curtis**	**Richness t**	**Richness KS**	**Shannon t**	**Shannon KS**
IBD	**MIDAS**	0.9993	1.000	0.6644	0.1829	0.2610	0.2637
	D-M	0.0090	<0.0001	0.3303	<0.0001	<0.0001	<0.0001
	MetaSPARSim	0.0340	<0.0001	0.3102	<0.0001	0.0078	<0.0001
	SparseDOSSA	0.7972	<0.0001	0.0569	<0.0001	<0.0001	<0.0001
MOMS-PI	**MIDAS**	0.5793	0.8617	0.6252	0.0019	<0.0001	<0.0001
	D-M	<0.0001	<0.0001	0.0028	<0.0001	<0.0001	<0.0001
	MetaSPARSim	<0.0001	<0.0001	0.6341	<0.0001	<0.0001	<0.0001
	SparseDOSSA	<0.0001	<0.0001	<0.0001	<0.0001	0.0002	0.0015

* P-values obtained from PERMANOVA.

template data, we used the Welch t-test to compare the means and the Kolmogorov-Smirnov two-sample test to compare the full distribution. In Table 1 we report the average p-value obtained from 20 simulated datasets for each method. In Figure 4, we also plot the alpha diversities for a single data set from each simulation method. For the IBD data, all methods successfully reproduced the mean richness, while for the MOMS-PI data only MIDAS and MetaSPARSim had mean richness that did not significantly differ from the template data. The situation is very different if we ask for the entire distribution of sample richness values to be the same as the template data; here, only MIDAS generated data that met this criterion when using the IBD data as the template. For the Shannon index, only MIDAS produced a non-significant t-test for the Shannon index for the IBD data, while none of the methods met this criterion for the MOMS-PI data. The same pattern was found for the full distribution of Shannon values.

We also compare the computational time that each method takes to fit its proposed model to the template IBD and MOMS-PI datasets and to simulate one dataset of the same size, which is summarized in Table 2. The computational time is evaluated on an Intel Quad core 2.7GHz processor, with 8GB memory. Comparing the total time used, MIDAS is one of the fastest, especially when fitting and simulating the large dataset of MOMS-PI. For the model fitting, MetaSPARSim is the fastest, but it is very slow in generating new data. For generating new data with specified parameters, D-M is the fastest. The computation time of SparseDOSSA for fitting the model depends on the number of iterations in its EM algorithm. In practice, it takes more than 3 hours to fit its

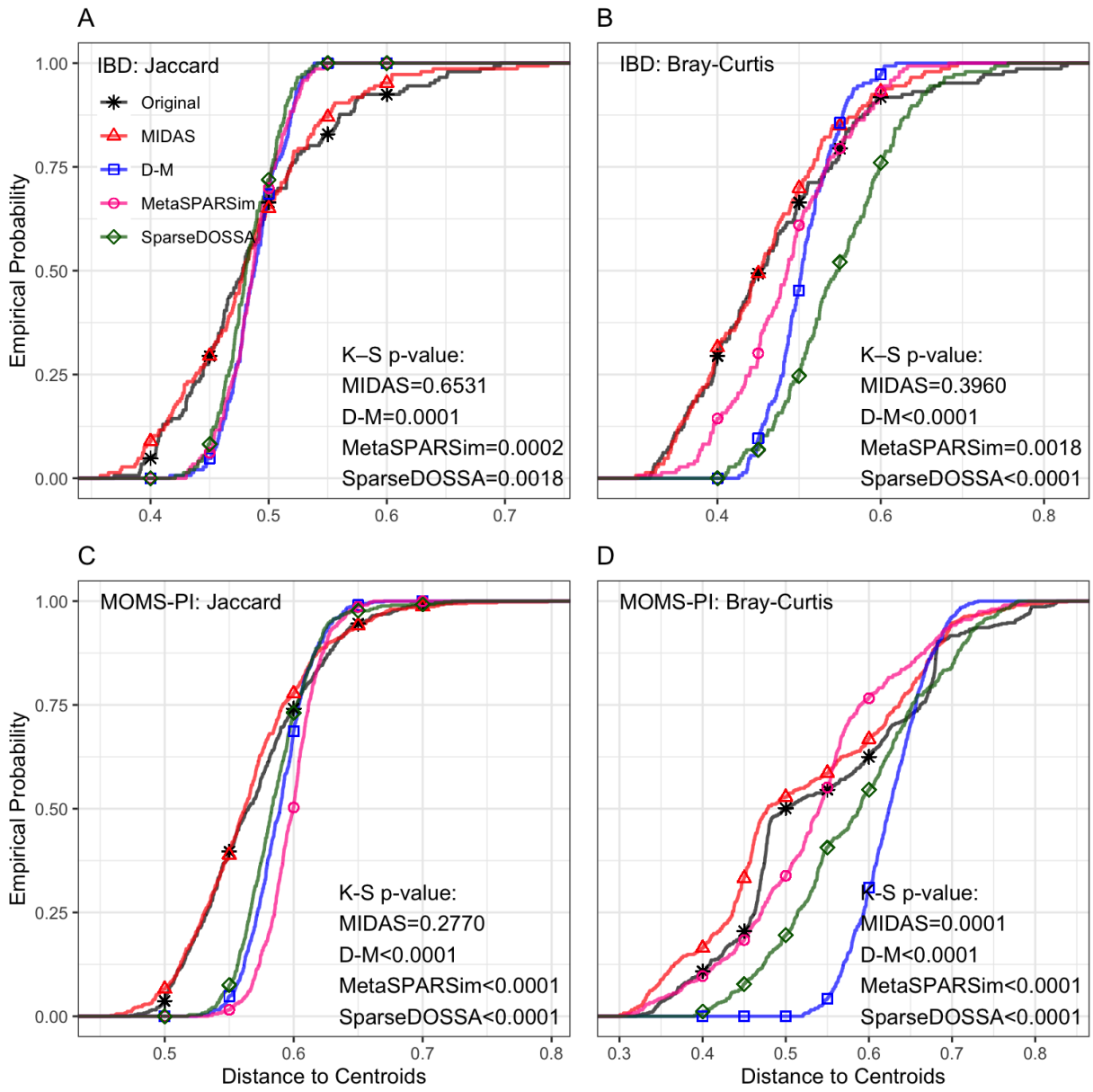

Figure 3: Empirical cumulative distribution function of distances to centroids

model to our IBD and MOMSPI datasets, making it hard to use in practice; the pre-trained models can be used if faster results are needed, but then a user-selected template dataset cannot be used. Discounting the time required for model fitting, MIDAS, D-M and SparseDOSSA all can generate replicate datasets quickly; MetaSPARSim is the only outlier in this regard.

Finally, in the supplementary material, we show the success of our approach to changing the library size for simulated data. Using rarefied data as the template, the data we simulated with the

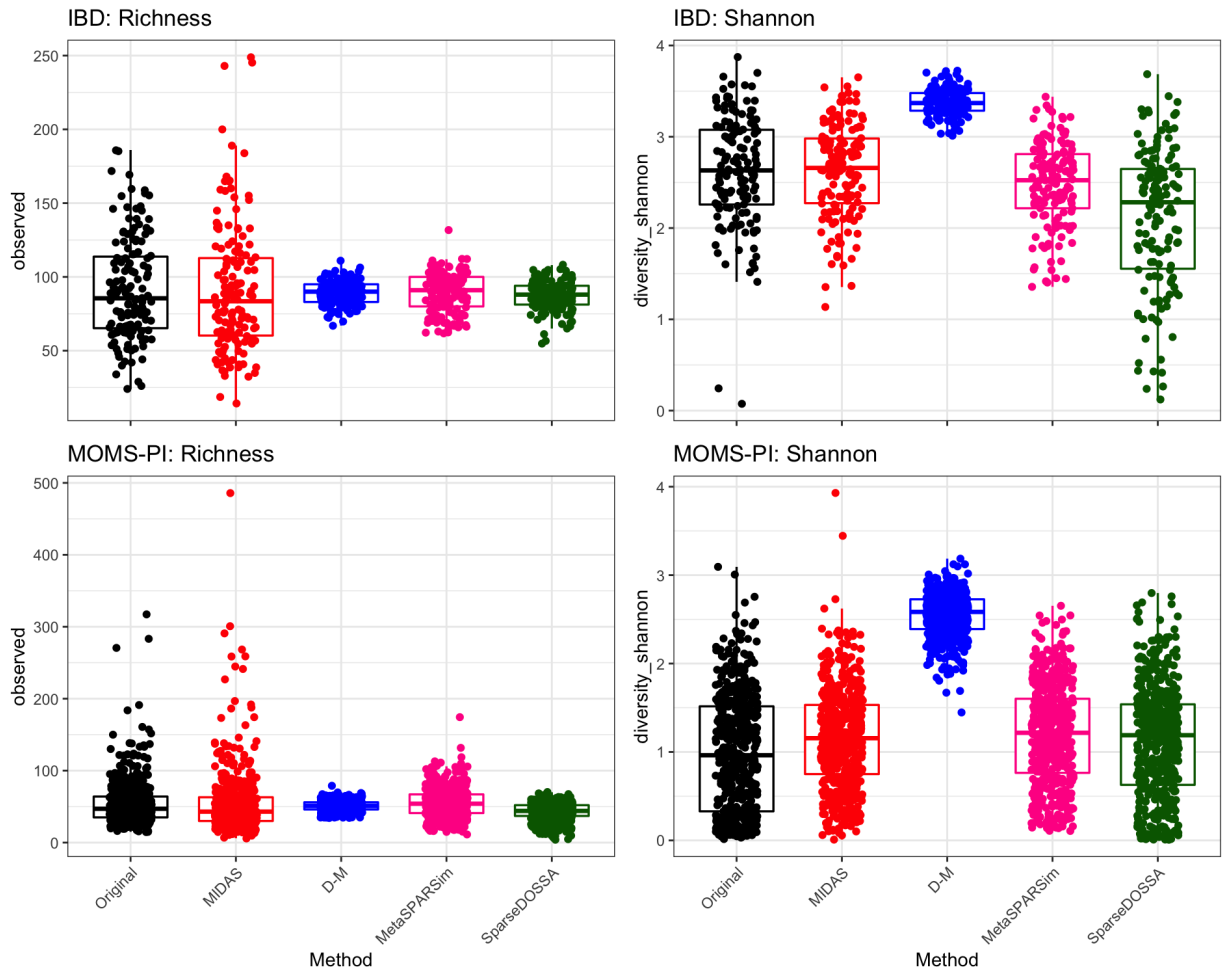

Figure 4: Alpha diversities (Richness and Shannon Index) of original and a single simulated dataset for each of four simulation methods.

original (unrarefied) library size agrees well with the original template data, both by ordination (Supplementary Figure S3) and in terms of beta dispersion (as measured by the distance from each observation to the data centroid) (Supplementary Figure S4).

Table 2: Computation time (seconds) required to fit the template data, and to simulate a new dataset with the same library size. Simulating time is the average time over 20 replicates of generating datasets of the same size as the real data. Total time is the sum of fitting and simulating times.

Method	**IBD**			**MOMS-PI**		
	Fitting	**Simulating**	**Total**	**Fitting**	**Simulating**	**Total**
MIDAS	25.5	2.5	28.0	162.0	15.3	177.6
D-M	25.0	0.3	25.3	308.4	2.2	310.6
MetaSPARSim	7.4	144.9	152.3	41.3	469.4	510.7
SparseDOSSA	10812.6	0.8	10813.4	11792.5	5.2	11797.7

4 Discussion

Simulating realistic microbiome datasets is essential for methodology development in microbiome studies. However, this task is surprisingly difficult due to the complexity of microbiomial abundance data. Here we propose a two-step MIcrobiome DAta Simulator (MIDAS) to generate microbiome datasets that resemble a real (template) microbiome dataset of the user's choice. The first step of MIDAS involves generating correlated binary indicators that represent the presence-absence status of all taxa. The correlation matrix for these binary indicators is estimated from the real data with a shrinkage procedure. The second step of MIDAS generates relative abundance and counts for the taxa that are considered to be present in step 1, utilizing a Gaussian copula to account for the taxon-taxon correlations.

In simulation studies, it is frequently necessary to generate microbiome data for which relative abundances or presence-absence status are related to a trait or covariate that are associated with one or a few clinical characteristics. MIDAS supports this situation; taxon relative abundances, library sizes, and taxon proportion of zero cells are output by the MIDAS set-up function `Midas.setup`. These quantities can be modified using the function `Midas.modify` before passing them as inputs

into the MIDAS simulation function `Midas.sim`. In the simple situation with case and control, it may be worthwhile to restrict the template data to unaffected controls.

We expect that modifying library sizes will perform best if the new library sizes are within the range of library sizes found in the template data, so that the number of zero cells for a sample can be reasonably predicted from its new library size. In our simulations, we found MIDAS could successfully execute an 11% increase in library size while remaining faithful to the template data.

MIDAS is designed to generate a new dataset that is as close as possible to the target data. In some cases, it may be desirable to introduce extra variability. For example, in section 2.1 we used the number of non-zeroes d_j found in the original data; we could instead sample these from a distribution, possibly while maintaining their rank order, or add random noise. In section 2.2, we generated relative abundance values $\widetilde{\pi}_{\cdot j}$ using a sampling-without-replacement scheme. We could instead sample from a smooth version of F_j (say, using a linear interpolation of the CDF) or even fit a Beta distribution to the non-zero relative abundances for each taxon. Finally, a taxon j that appears in all samples will always appear in all simulated samples; this could easily be modified by allowing a zero cell in the ith sample with probability ε_{1j}. Similarly, we could allow a taxon j' that have no counts in any sample to appear in a simulated sample with probability $\varepsilon_{0j'}$; we could provide a distribution of relative abundances for such a taxon by using the distribution of relative abundances from a randomly-selected taxon having the same number of observed non-zero cells as taxon j' has. We may consider these changes as we receive feedback from users.

To summarize, MIDAS is easy to implement, flexible and suitable for most microbiome data simulation situations. For the two template datasets we considered, MIDAS showed superior performance when compared to existing competitors both by PERMANOVA and in terms of beta dispersion.

Data Availability

All template datasets are publicly available and can be accessed through R package `HMP2Data`. Details can be found in the vignette of R package `MIDAS`.

Competing interests

The authors declare no competing risk.

Acknowledgements

Funding

Drs Zhao and Satten's work is supported, in part, by National Institute of Health (R01GM147162),

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. Austral Ecology, **26**.

Bonnet, D. G. and Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. Journal of Educational and Behavioral Statistics, **30(2)**, 213–225.

Chen, J. and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. Annals of Applied Statistics, **7**.

Dejea, C. M., Wick, E. C., Hechenbleikner, E. M., White, J. R., Mark Welch, J. L., Rossetti, B. J., Peterson, S. N., Snesrud, E. C., Borisy, G. G., Lazarev, M., Stein, E., Vadivelu, J., Roslani, A. C., Malik, A. A., Wanyiri, J. W., Goh, K. L., Thevambiga, I., Fu, K., Wan, F., Llosa, N., Housseau, F., Romans, K., Wu, X., McAllister, F. M., Wu, S., Vogelstein, B., Kinzler, K. W.,

Pardoll, D. M., and Sears, C. L. (2014). Microbiota organization is a distinct feature of proximal colorectal cancers. Proc. Natl. Acad. Sci. U.S.A., **111**(51), 18321–18326.

Dixon, P. (2003). Vegan, a package of r functions for community ecology. Journal of Vegetation Science, **14**.

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., Huang, B., Arodz, T. J., Edupuganti, L., Glascock, A. L., Xu, J., Jimenez, N. R., Vivadelli, S. C., Fong, S. S., Sheth, N. U., Jean, S., Lee, V., Bokhari, Y. A., Lara, A. M., Mistry, S. D., Duckworth, R. A., Bradley, S. P., Koparde, V. N., Orenda, X. V., Milton, S. H., Rozycki, S. K., Matveyev, A. V., Wright, M. L., Huzurbazar, S. V., Jackson, E. M., Smirnova, E., Korlach, J., Tsai, Y. C., Dickinson, M. R., Brooks, J. L., Drake, J. I., Chaffin, D. O., Sexton, A. L., Gravett, M. G., Rubens, C. E., Wijesooriya, N. R., Hendricks-Muñoz, K. D., Jefferson, K. K., Strauss, J. F., and Buck, G. A. (2019). The vaginal microbiome and preterm birth. Nature Medicine, **25**.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. Nature Medicine, **24**.

Hu, Y., Satten, G. A., and Hu, Y.-J. (2022). Locom: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. Proceedings of the National Academy of Sciences.

Hu, Y. J. and Satten, G. A. (2020). Testing hypotheses about the microbiome using the linear decomposition model (ldm). Bioinformatics, **36**.

Hu, Y. J., Lane, A., and Satten, G. A. (2021). A rarefaction-based extension of the ldm for testing presence-absence associations in the microbiome. Bioinformatics, **37**.

Jiang, Z., He, M., Chen, J., Zhao, N., and Zhan, X. (2022). MiRKAT-MC: A Distance-Based Microbiome Kernel Association Test With Multi-Categorical Outcomes. Front Genet, **13**, 841764.

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R. A., Ogino, S., Birren, B. W., Huttenhower, C., Garrett, W. S., and Meyerson, M. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res., **22**(2), 292–298.

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

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, R. A., Bishai, J., Bullock, K., Deik, A., Dennis, C., Kaplan, J. L., Khalili, H., McIver, L. J., Moran, C. J., Nguyen, L., Pierce, K. A., Schwager, R., Sirota-Madi, A., Stevens, B. W., Tan, W., ten Hoeve, J. J., Weingart, G., Wilson, R. G., Yajnik, V., Braun, J., Denson, L. A., Jansson, J. K., Knight, R., Kugathasan, S., McGovern, D. P., Petrosino, J. F., Stappenbeck, T. S., Winter, H. S., Clish, C. B., Franzosa, E. A., Vlamakis, H., Xavier, R. J., and Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, **569**.

Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., Tickle, T. L., Lu, Y., Carmody, R. N., Franzosa, E. A., Janson, L., and Huttenhower, C. (2021). A statistical model for describing and simulating microbial community profiles. PLoS Computational Biology, **17**.

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

Martin, B. D., Witten, D., and Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. Ann Appl Stat, **14**(1), 94–115.

Patuzzi, I., Baruzzo, G., Losasso, C., Ricci, A., and Camillo, B. D. (2019). Metasparsim: A 16s rrna gene sequencing count data simulator. BMC Bioinformatics, **20**.

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

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., and Huttenhower, C. (2019). The integrative human microbiome project. Nature, **569**.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. Statistics and Computing, **25**, 543–559.

Revelle, W. (2015). Package 'psych' - procedures for psychological, psychometric and personality research. R Package.

Ruppert, D. and Mattesson, D. S. (2015). Statistics and Data Analysis for Financial Engineering, with R examples. Springer, New York, NY.

Simren, M., Barbara, G., Flint, H. J., Spiegel, B. M., Spiller, R. C., Vanner, S., Verdu, E. F., Whorwell, P. J., and Zoetendal, E. G. (2013). Intestinal microbiota in functional bowel disorders: a Rome foundation report. Gut, **62**(1), 159–176.

Sze, M. A. and Schloss, P. D. (2016). Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. MBio, **7**(4).

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. Genome Medicine, **8**.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. American Journal of Human Genetics, **96**.