The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1476–1499
https://doi.org/10.1214/21-AOAS1552

DIRICHLET-TREE MULTINOMIAL MIXTURES FOR CLUSTERING MICROBIOME COMPOSITIONS

BY JIALIANG MAO[a] AND LI MA[b]

Department of Statistical Science, Duke University, [a]*jialiang.mao@duke.com,* [b]*li.ma@duke.edu*

Studying the human microbiome has gained substantial interest in recent years, and a common task in the analysis of these data is to cluster microbiome compositions into subtypes. This subdivision of samples into subgroups serves as an intermediary step in achieving personalized diagnosis and treatment. In applying existing clustering methods to modern microbiome studies, including the American Gut Project (AGP) data, we found that this seemingly standard task, however, is very challenging in the microbiome composition context, due to several key features of such data. Standard distance-based clustering algorithms generally do not produce reliable results, as they do not take into account the heterogeneity of the cross-sample variability among the bacterial taxa, while existing model-based approaches do not allow sufficient flexibility for the identification of complex within-cluster variation from cross-cluster variation. Direct applications of such methods generally lead to overly dispersed clusters in the AGP data, and such a phenomenon is common for other microbiome data. To overcome these challenges, we introduce Dirichlet-tree multinomial mixtures (DTMM) as a Bayesian generative model for clustering amplicon sequencing data in microbiome studies. DTMM models the microbiome population with a mixture of Dirichlet-tree kernels that utilizes the phylogenetic tree to offer a more flexible covariance structure in characterizing within-cluster variation, and it provides a means for identifying a subset of signature taxa that distinguish the clusters. We perform extensive simulation studies to evaluate the performance of DTMM, compare it to state-of-the-art model-based and distance-based clustering methods in the microbiome context and carry out a validation study on a publicly available longitudinal data set to confirm the biological relevance of the clusters. Finally, we report a case study on the fecal data from the AGP to identify compositional clusters among individuals with inflammatory bowel disease and diabetes. Among our most interesting findings is that enterotypes (i.e., gut microbiome clusters) are not always defined by the most dominant taxa, as previous analyses had assumed, but can involve a number of less abundant taxa which cannot be identified with existing distance-based and method-based approaches.

1. Introduction. The human microbiome is the collective genomes of all microbes that inhabit the human body. It has been associated with various aspects of our physiology (Karlsson et al. (2013), Qin et al. (2012), Turnbaugh et al. (2009)) and is suggested as a way toward precision medicine (Kuntz and Gilbert (2017)). The development of next-generation sequencing strategies enables us to profile the microbiome fast and economically through either amplicon sequencing on target genes (usually the 16S ribosomal RNA gene) or shotgun sequencing on the entire microbial genome. In this work we focus on datasets obtained from amplicon sequencing studies. Traditionally, the sequencing reads are sent to preprocessing pipelines, such as QIIME (Caporaso et al. (2010)), to construct clusters named operational taxonomic units (OTUs), based on certain predefined similarity threshold (typically 97%).

Received October 2020.
Key words and phrases. Bayesian hierarchical models, compositional data, latent variable models, probabilistic learning.

In contrast, more recently developed pipelines, such as DADA2 (Callahan et al. (2016)), directly resolve amplicon sequence variants (ASVs) which is shown to outperform OTUs in terms of accuracy and interpretability (Callahan, McMurdie and Holmes (2017)). OTUs and ASVs serve as the unit for downstream statistical analyses and provide the same interface: each sample is a vector of counts on a list of units (OTUs or ASVs), representing the composition of the underlying community. The methodology developed in this work applies to both OTUs and ASVs; we thus use the customary OTU to refer to the unit.

Given the heterogeneous nature of microbiome samples, a useful idea in microbiome analysis is to first group individual samples into clusters and seek to understand the relation of these clusters with the host environment and other health outcomes. For example, in the context of the human gut microbiome, these clusters are referred to as "enterotypes" (Arumugam et al. (2011), Costea et al. (2018)) which are shown to be associated with long-term dietary habits and risks for obesity and Crohn's disease (Holmes, Harris and Quince (2012), Quince et al. (2013), Wu et al. (2011)).

In applying this strategy to analyze several modern microbiome data sets, including the American Gut Project (AGP) (McDonald, Birmingham and Knight (2015), McDonald et al. (2018)), however, we found that reliably clustering microbiome samples is, in fact, very challenging. Off-the-shelf clustering algorithms, such as k-means, Partitioning Around Medoids (PAM) and hierarchical clustering that are distance-based, are not satisfactory when applied to microbiome data. This is true even when using popular distance metrics specifically tailored for microbiome compositions, such as the Bray–Curtis dissimilarity and the Unifrac distances (Lozupone and Knight (2005)). Other authors, including Koren et al. (2013), also showed that, in clustering microbiome compositions, different methods for selecting the number of clusters in distance-based methods can yield inconsistent results and that these algorithms are highly sensitive to the distance metrics chosen. We believe a main reason for such inconsistency is that different distance metrics induce different weighting on the OTUs that are inconsistent with the actual patterns of heterogeneous cross-sample variability in the data.

In addition, we have found that existing nondistance-based methods, such as those based on probabilistic models like mixtures, also suffer similar issues for microbiome compositional data. In particular, the arguably most widely used model-based approach for clustering microbiome data, called the Dirichlet multinomial mixture (DMM) model (Holmes, Harris and Quince (2012)), which adopts a multinomial sampling scheme and generates the sample-specific multinomial parameters from a finite mixture of Dirichlet components, often results in very large and overly dispersed clusters. A key reason for this phenomenon, we believe, is the model's lack of flexibility in characterizing the cross-sample variability which, in turn, hampers the ability to differentiate within-cluster variability from between-cluster variability. In particular, the single dispersion (also called concentration) parameter of the Dirichlet distribution is insufficient for characterizing the often complex variation among samples within each cluster. Moreover, the Dirichlet distribution implies a priori independence among OTU compositions, up to the sum to one constraint (Aitchison (1982)), which is restrictive in the microbiome context (Wang and Zhao (2017)).

To overcome these difficulties and achieve more reliable cluster analysis on the AGP study and other microbiome data, we introduce a new probabilitistic model for clustering microbiome compositions, called Dirichlet-tree multinomial mixtures (DTMM). Similar to the DMM, our method uses mixture modeling to achieve clustering under the Bayesian inference framework. The difference is twofold. First, by utilizing a natural hierarchical relationship among the OTUs in terms of the phylogenetic tree, DTMM adopts the Dirichlet-tree distribution (DT) (Dennis (1991), Wang and Zhao (2017)) as the mixture component, in contrast to the Dirichlet component used by DMM. The DT mixing component incorporates multiple dispersion parameters, one for each node in the phylogenetic tree, thereby allowing more

flexible and realistic cross-sample variation among the OTUs. In addition, motivated by the fact that microbiome clusters are often determined by a subset of the taxa, we incorporate a model selection feature into the DTMM framework that allows: (i) the signature taxa that distinguish the clusters to be identified and (ii) the common features across clusters (e.g., groups of "house-keeping" taxa) to be more accurately characterized through borrowing information among clusters.

With the proposed method we report a case study of the AGP data to find and explore enterotypes of samples that are diagnosed with inflammatory bowel disease (IBD) or diabetes. Enterotypes have been established and compared among samples from different geographical locations (Arumugam et al. (2011), Costea et al. (2018)) or with different host dietary patterns (Wu et al. (2011)). Our analysis provides another important facet to this thread of work. Both IBD and diabetes were shown to be related to the human gut microbiome (Kostic, Xavier and Gevers (2014), Qin et al. (2012)). It is thus natural to expect different enterotype patterns in samples with these diseases. Interestingly and contrary to traditional wisdom, our analysis shows that enterotypes are not always characterized by a small number of highly abundant dominant taxa but can arise from combinations of several taxa.

The rest of the paper is organized as follows. In Section 2 we introduce a phylogenetic tree-based decomposition of the multinomial counts and proposes the DTMM model for clustering OTU counts based on this decomposition. In Section 3 we conduct a series of representative numerical experiments to evaluate the performance of DTMM. We also validate the clusters found by DTMM with a longitudinal microbiome dataset for which the actual clustering pattern is roughly known, due to the experiment design. In Section 4 we report our case study of the AGP data. Section 5 concludes with a few remarks.

2. Method.

2.1. *DM and DMM.* Consider a microbiome dataset with OTU counts of n samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$. Each sample is a vector of counts of the M OTUs in the study denoted by $\Omega = \{\mathrm{OTU}_2, \mathrm{OTU}_2, \ldots, \mathrm{OTU}_M\} = \{\omega_1, \omega_2, \ldots, \omega_M\}$. Let the ith sample and the counts in that sample be $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iM})$ and $N_i = \sum_{j=1}^{M} y_{ij}$, where y_{ij} is the count of OTU j. The samples can be stacked into an OTU table, denoted by $\boldsymbol{Y}$, as shown in Table 1. In this work we treat the total counts N_i's as given since they are artificial quantities that depend on the sequencing depth. We consider the following Dirichlet-multinomial model (DM) (Knights et al. (2011), La Rosa et al. (2012)):

$$\boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i \overset{\text{ind}}{\sim} \mathrm{Mult}(N_i, \boldsymbol{p}_i) \quad \text{and} \quad \boldsymbol{p}_i \mid \boldsymbol{\alpha} \overset{\text{iid}}{\sim} \mathrm{Dir}(\boldsymbol{\alpha}), \tag{1}$$

where $\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iM})$, p_{ij} is the probability that a count in sample i belongs to OTU j, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ with $\alpha_j > 0$ for $j = 1, \ldots, M$.

Viewing each sample as randomly drawn from an underlying community characterized by its multinomial parameter (Holmes, Harris and Quince (2012)), DM models all the communities as realizations of a single metacommunity governed by $\boldsymbol{\alpha}$. Holmes, Harris and Quince

TABLE 1
An $n \times M$ OTU table

Sample	ω_1	ω_2	$\cdots$	ω_M	Sum
1	y_{11}	y_{12}	$\cdots$	y_{1M}	N_1
2	y_{21}	y_{22}	$\cdots$	y_{2M}	N_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nM}	N_n

(2012) extend DM to Dirichlet multinomial mixtures (DMM) by replacing the single Dirichlet prior in DM with a finite mixture of K Dirichlets,

$$\boldsymbol{p}_i \mid \boldsymbol{\pi}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \operatorname{Dir}(\boldsymbol{\alpha}_k) \quad \text{and} \quad \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \operatorname{Dir}(\boldsymbol{b}_0), \tag{2}$$

where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kM})$, $\boldsymbol{\pi}$ the weights of the metacommunities with $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$ for $k = 1, \ldots, K$. In DMM each sample is viewed as a draw from a unique community that is itself drawn from one of the K metacommunities.

As a clustering method, DMM has several limitations. Most importantly, it could not adequately model the within-cluster variation of the microbial composition. This can be seen by writing the cluster-specific Dirichlet parameter $\boldsymbol{\alpha}_k$ as $\boldsymbol{\alpha}_k = \alpha_{k0} \cdot \bar{\boldsymbol{\alpha}}_k$, where $\bar{\boldsymbol{\alpha}}_k$, lying in the $(M-1)$-dimensional simplex, represents the prior mean of the multinomial probabilities in cluster k; $\alpha_{k0} = \sum_{j=1}^{M} \alpha_j$ determines the within-cluster variation of all these probabilities around $\bar{\boldsymbol{\alpha}}_k$ simultaneously. Second, the multinomial parameters in DM are modeled independently, up to the sum to one constraint (Mosimann (1962)), which is not suitable in the microbiome context since the OTUs are functionally and evolutionarily related. Although DMM is specified within the Bayesian framework, the posterior inference is performed by optimization through an EM algorithm with Laplace approximations of the marginal likelihoods. When the number of OTUs is moderate, which is typical in microbiome studies, these techniques are numerically unstable and cannot provide reliable uncertainty quantifications.

2.2. *Dirichlet-tree multinomial mixtures.* OTUs in a microbiome study are evolutionarily related. Typically, this relationship can be summarized into a rooted phylogenetic tree, where each internal node can be viewed as a "taxa" that represents the most recent common ancestor of its descendant OTUs. Let $\mathcal{T} = \mathcal{T}(\mathcal{I}, \mathcal{U}; \mathcal{E})$ be a rooted full binary phylogenetic tree over the M OTUs in the study, where $\mathcal{I}$, $\mathcal{U}$ and $\mathcal{E}$ denote the set of internal nodes, leaves and edges of $\mathcal{T}$, respectively. We denote each node $A \in \mathcal{I} \cup \mathcal{U}$ by the set of its descendant OTUs. In particular, $A = \Omega$ denotes the root of $\mathcal{T}$; $A = \{\omega_j\}$ represents the leaf that contains OTU j for $j = 1, \ldots, M$. With our notation, $\mathcal{U} = \{\{\omega\} : \omega \in \Omega\}$. For $A \in \mathcal{I}$, let A_l and A_r be the left and right children of A, respectively. For $A \in \mathcal{I} \cup \mathcal{U} \setminus \{\Omega\}$, let A_p be its parent and A_s be its sibling (i.e., the node in $\mathcal{T}$ that has the same parent as A).

Given $\mathcal{T}$, it can be shown that the multinomial likelihood of $\boldsymbol{y}_i$ factorizes into a series of binomial likelihoods at the internal nodes of $\mathcal{T}$,

$$\mathcal{L}_M(\boldsymbol{y}_i \mid \boldsymbol{p}_i) \propto \prod_{\{A : A \in \mathcal{I}\}} \mathcal{L}_B\big(y_i(A_l) \mid y_i(A), \theta_i(A)\big), \tag{3}$$

where

$$y_i(A) = \sum_{\{j : \omega_j \in A\}} y_{ij}, \qquad \theta_i(A) = \frac{\sum_{\{j : \omega_j \in A_l\}} p_{ij}}{\sum_{\{j : \omega_j \in A\}} p_{ij}}, \tag{4}$$

$$\text{and} \qquad y_i(A_l) \mid y_i(A), \theta_i(A) \overset{\text{ind}}{\sim} \operatorname{Binom}\big(y_i(A), \theta_i(A)\big).$$

Note that, for $j = 1, \ldots, M$, there is a unique path $\mathscr{P}^j = A_0^j = \Omega \to A_1^j \to \cdots \to A_{l_j}^j \to \{\omega_j\}$ in $\mathcal{T}$ connecting the root with $\{\omega_j\}$ such that

$$p_{ij} = \prod_{l=0}^{l_j} \theta_i(A_l^j). \tag{5}$$

We denote $\boldsymbol{\theta}_i = \{\theta_i(A) : A \in \mathcal{I}\}$. Let $\boldsymbol{\theta}_i = \operatorname{tr}(\boldsymbol{p}_i)$ and $\boldsymbol{p}_i = \operatorname{tr}^{-1}(\boldsymbol{\theta}_i)$ be the "tree-based ratio transform" and the "inverse tree-based ratio transform" defined in (4) and (5). $\boldsymbol{p}_i$ and $\boldsymbol{\theta}_i$ give

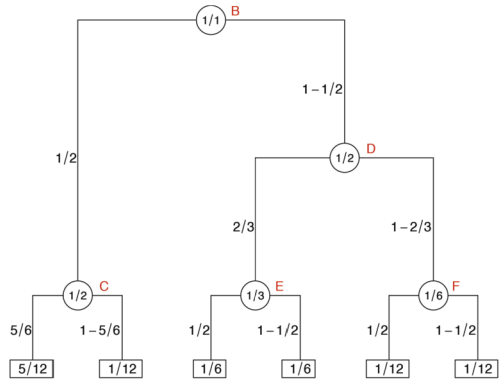


FIG. 1. *A tree-based generation of* $\boldsymbol{p}_{ex}$.

two equivalent parameterizations of the distribution of $\boldsymbol{y}_i$. Figure 1 gives an example of how a specific multinomial parameter $\boldsymbol{p}_{\text{ex}} = (\frac{5}{12}, \frac{5}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12})$ on *six* OTUs can be generated sequentially along a given tree.

The likelihood factorization in (3) provides an orthogonal decomposition of the empirical evidence about $\boldsymbol{p}_i$ into pieces of evidence about $\theta_i(A)$ at $A \in \mathcal{I}$ which suggests a divide-and-conquer strategy of doing inference on $\boldsymbol{p}_i$ through learning the branching probability $\boldsymbol{\theta}_i$. To this end, we model the binomial parameters with independent beta variables,

$$\theta_i(A) \mid \theta(A), \tau(A) \overset{\text{ind}}{\sim} \text{Beta}\big(\theta(A)\tau(A), \big(1-\theta(A)\big)\tau(A)\big), \text{ for all } A \text{ and } i, \tag{6}$$

where $\theta(A) \in (0, 1)$ is the mean of $\theta_i(A)$ and $\tau(A) > 0$ is a dispersion parameter that controls the variability of $\theta_i(A)$ around its mean. Note that the independent betas on $\boldsymbol{\theta}_i$, together with the relation in (5), induce a joint model on $\boldsymbol{p}_i$ which falls into the family of Dirichlet-tree distributions (DT) (Dennis (1991)). Let $\boldsymbol{\theta} = \{\theta(A) : A \in \mathcal{I}\}$ and $\boldsymbol{\tau} = \{\tau(A) : A \in \mathcal{I}\}$; we shall denote the Dirichlet-tree model on $\boldsymbol{p}_i$ as

$$\boldsymbol{p}_i = \text{tr}^{-1}(\boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i \sim \text{DT}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\tau}). \tag{7}$$

When $\tau(A) = \tau(A_l) + \tau(A_r)$ for every $A \in \mathcal{I}$ that has nonleaf children, DT degenerates to the Dirichlet distribution. Without this constraint, DT offers a more flexible way to model the variability of $\boldsymbol{p}_i$ around its cluster centroid.

DT also induces a more flexible covariance structure than the Dirichlet distribution. This can be seen by considering the covariance between any two different categories j_1 and j_2. Suppose that the first $(L+1)$ nodes in $\mathcal{P}^{j_1}$ and $\mathcal{P}^{j_2}$ are shared, and let the shared path be $\Omega = A_0 \to A_1 \to \cdots \to A_L$. It can be shown that (Dennis (1991))

$$\text{Cov}(p_{ij_1}, p_{ij_2}) = \left[\frac{\tau(A_L)}{\tau(A_L)+1} \prod_{1 \le t \le L} \frac{[a(A_t)+1]\tau(A_{t-1})}{a(A_t)[\tau(A_{t-1})+1]} - 1\right] \mathbb{E}(p_{ij_1})\mathbb{E}(p_{ij_2}), \tag{8}$$

where $a(A_t) = \theta(A_{t-1})\tau(A_{t-1})$ if A_t is the left child of A_{t-1} and $a(A_t) = (1-\theta(A_{t-1})) \times \tau(A_{t-1})$ otherwise. In DT the covariance between categories depends not only on their means and the sum of the pseudo counts as in the Dirichlet distribution but also on the tree structure. This offers a more flexible covariance structure among OTU counts governed by the phylogenetic information. For example, since $a(A_t) < \tau(A_{t-1})$, $[a(A_t)+1]\tau(A_{t-1})/a(A_t)[\tau(A_{t-1})+1] > 1$, p_{ij_1} and p_{ij_2} can be positively correlated if j_1 and j_2 share a series of common ancestors in the phylogenetic tree. On the other hand, if j_1 and j_2 are far away in the phylogenetic tree such that their only common ancestor is Ω, $\text{Cov}(p_{ij_1}, p_{ij_2}) = -\mathbb{E}(p_{ij_1})\mathbb{E}(p_{ij_2})/(\tau(\Omega)+1)$, as in the Dirichlet distribution. When the

phylogenetic tree gives decent summaries of the functional relationship among OTUs, this introduces suitable covariance structure among the OTU counts and can improve the inference substantially.

DT has been used for microbiome modelings in various contexts for different purposes. For example, Wang and Zhao (2017) apply the DT multinomial model to study the association between OTU counts and a set of covariates; Tang, Ma and Nicolae (2018) and Mao, Chen and Ma (2020) use the tree decomposition to motivate a divide-and-conquer strategy to increase the statistical power when comparing the OTU composition of groups of samples.

In this work we replace the Dirichlet component in DMM with DT mixing components to give a more suitable clustering model for microbiome data. In DMM, if the counts of one or a small number of OTUs are highly variable, the single dispersion parameter would be estimated large in adjustment of this variation. As a result, less variable cluster signatures contained in other OTUs would be buried, and the samples would be modeled as drawn from a single or otherwise small number of highly heterogeneous clusters. In contrast, the set of dispersion parameters in DT are able to account for different levels of variation across OTUs and thus prevent the signals from being contaminated by the noises.

Specifically, we take the multinomial sampling scheme as in DMM. Following (7), let

$$\boldsymbol{\theta}_i \mid \boldsymbol{\pi}, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^K \overset{\text{iid}}{\sim} \sum_{k=1}^K \pi_k \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \quad \text{and} \quad \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\boldsymbol{b}_0), \tag{9}$$

where $(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \overset{\text{iid}}{\sim} \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_0, \boldsymbol{\nu}_0) \times F(\boldsymbol{\tau})$, $\mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_0; \boldsymbol{\nu}_0)$ the population model for the cluster centroids, $F(\boldsymbol{\tau}) = \prod_{A \in \mathcal{I}} F^A(\tau(A))$ the model for the within-cluster dispersion. Note that $(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)$ determines the kth meta-community. We shall for now refer to this model as the Dirichlet-tree multinomial mixtures (DTMM).

2.3. *Discriminative taxa selection*. In DMM all OTUs are treated equally in the clustering procedure. In many applications, however, it is expected that only a (possibly small) subset of OTUs determine the underlying clusters. When this is the case, not only can identifying these signature taxa enhance the sensitivity for separating the clusters, but it will also improve the interpretability of the resulting inference.

In this section we incorporate automatic taxa selection into DTMM. At the same time, in specifying the DTMM model we adopt a standard nonparametric modeling approach in dealing with the difficulty in setting the number of clusters beforehand by replacing the finite mixture with a Dirichlet process (DP) mixture. Formally, for $A \in \mathcal{I}$, let $\gamma(A) \in \{0, 1\}$ be an indicator of whether node A can be contributive to the latent clustering: $\gamma(A) = 1$, if A can play a role in defining clusters, and 0 otherwise. If $\gamma(A) = 1$, A is "active" in clustering, and we allow different clusters to have cluster-specific branching probabilities at A; otherwise, A is "inactive," and we force all the clusters at A to share the same branching probability. For this reason we shall refer to $\gamma(A)$ and $\lambda(A)$ as the activation indicator and the prior activation probability on A. Let $\boldsymbol{\gamma} = \{\gamma(A) : A \in \mathcal{I}\}$ be the collection of activation indicators of all the internal nodes.

Let $F(\cdot)$ be a probability measure on $(0, \infty)$ and $\delta_x(\cdot)$ the Dirac measure. The model can be written in the following hierarchical form:

- sampling model on $\boldsymbol{y}_i$:

$$\boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i \overset{\text{ind}}{\sim} \mathrm{Mult}(N_i, \boldsymbol{p}_i); \tag{10}$$

- model for the sample-specific compositional probability vector $\boldsymbol{p}_i$:

$$\boldsymbol{p}_i = \mathrm{tr}^{-1}(\boldsymbol{\theta}_i) \quad \text{and} \quad \boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i', \boldsymbol{\tau}_i' \overset{\text{ind}}{\sim} \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i'); \tag{11}$$

- model for the cluster-specific branching probabilities:

$$(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i') \mid G \overset{\text{iid}}{\sim} G \quad \text{and} \quad G \sim \mathrm{DP}\big(G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}); \beta\big); \tag{12}$$

- the base measure in DP:

$$\begin{aligned} &G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}) \\ &\quad = \prod_{A \in \mathcal{I}} G_0^A\big(\theta(A), \tau(A) \mid \gamma(A), \tilde{\theta}(A), \tilde{\tau}(A)\big), \\ &G_0^A\big(\theta(A), \tau(A) \mid \gamma(A) = 1, \tilde{\theta}(A), \tilde{\tau}(A)\big) \\ &\quad = \mathrm{Beta}\big(\theta_0(A)\nu_0(A), \big(1 - \theta_0(A)\big)\nu_0(A)\big) \times F^A(\tau), \\ &G_0^A\big(\theta(A), \tau(A) \mid \gamma(A) = 0, \tilde{\theta}(A), \tilde{\tau}(A)\big) \\ &\quad = \delta_{(\tilde{\theta}(A), \tilde{\tau}(A))}; \end{aligned} \tag{13}$$

- priors for the parameters in the base measure: for $A \in \mathcal{I}$,

$$\begin{aligned} \gamma(A) &\overset{\text{ind}}{\sim} \mathrm{Binom}\big(\lambda(A)\big), \\ \big(\tilde{\theta}(A), \tilde{\tau}(A)\big) &\overset{\text{ind}}{\sim} \mathrm{Beta}\big(\theta_0(A)\nu_0(A), \big(1 - \theta_0(A)\big)\nu_0(A)\big) \times F^A(\tau), \\ \lambda(A) &\overset{\text{ind}}{\sim} \mathrm{Beta}\big(a_0(A), b_0(A)\big). \end{aligned} \tag{14}$$

For simplicity, we shall refer to this model with taxa selection and infinite mixture still simply as the Dirichlet-tree multinomial mixtures (DTMM). The graphical model representation of DTMM is shown in Figure 2. Note that G in (12) is supported on a countable number of values since samples from a Dirichlet process are discrete, implying ties in the i.i.d. samples $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s and thus a clustering on i. This becomes clear with the stick-breaking construction

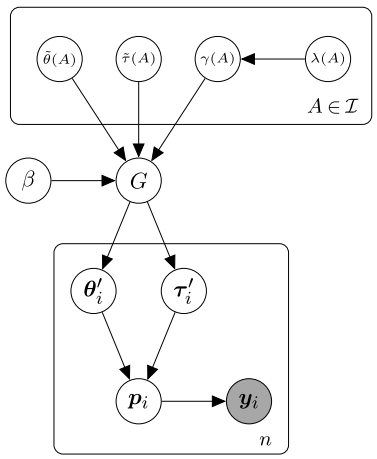


FIG. 2. *A graphical model representation of DTMM.*

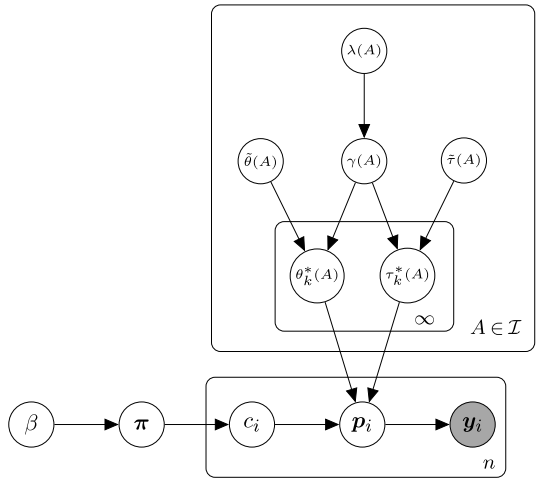


FIG. 3. *An alternative graphical model representation of DTMM.*

of the Dirichlet process (Sethuraman (1994)) from which we can rewrite (11) and (12) as

$$
\begin{aligned}
&\boldsymbol{p}_i \mid \boldsymbol{\pi}, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^{\infty} \overset{\text{iid}}{\sim} \sum_{k=1}^{\infty} \pi_k \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*), \\
&\qquad \pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \quad \text{where } v_1, v_2, \ldots \mid \beta \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \beta), \\
&(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*) \overset{\text{iid}}{\sim} G_0(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}}).
\end{aligned} \tag{15}
$$

For $i = 1, \ldots, n$, let $c_i \in \mathbb{N}^+$ be the cluster label for the ith sample such that $\boldsymbol{p}_i \mid c_i, \{(\boldsymbol{\theta}_k^*, \boldsymbol{\tau}_k^*)\}_{k=1}^{\infty} \sim \mathrm{DT}_{\mathcal{T}}(\boldsymbol{\theta}_{c_i}^*, \boldsymbol{\tau}_{c_i}^*)$. We can equivalently illustrate DTMM, as in Figure 3. For comparison, we can introduce the latent cluster labels to DMM and DTMM without taxa selection and write their graphical model representations, as in Figure 4 and Figure 5. Figure 5 and Figure 3 illustrate how DTMM is generalized in this section.

Prior specification. To complete the model specification, we need to choose $a_0(A)$, $b_0(A)$, $\theta_0(A)$, $\nu_0(A)$ and $F^A(\tau)$ for each $A \in \mathcal{I}$. Ideally, informative prior knowledge shall be incorporated in choosing these parameters. If, instead, no prior knowledge is available, we treat these parameters (priors) as global such that they do not depend on A and remove the "(A)" from the notation.

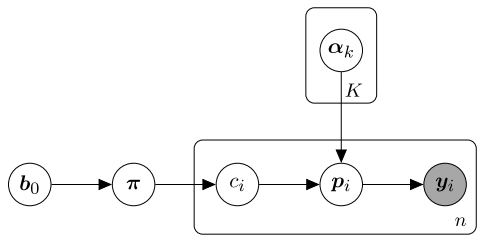


FIG. 4. *A graphical model representation of DMM.*

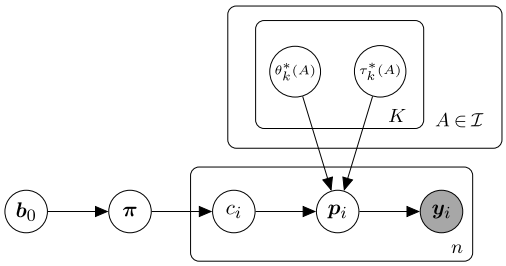


FIG. 5. *A graphical model representation of DTMM without taxa selection.*

For λ we could set $a_0 = b_0 = 1$ such that λ has a uniform distribution a priori, which yields the following prior probability on the $\gamma(A)$'s (Scott and Berger (2010)):

$$\Pr(\boldsymbol{\gamma}) = \frac{1}{M}\binom{M-1}{\sum_{A\in\mathcal{I}}\gamma(A)}^{-1}. \tag{16}$$

This prior allows multiplicity adjustment in the taxa selection. A default choice for (θ_0, ν_0) is $(0.5, 1)$ which yields the Jeffrey's prior on $\theta_k^*(A)$ and $\theta^*(A)$. For $F(\gamma)$, any prior with a reasonably large support that covers a wide range of dispersion levels can be chosen. For example, we let $F(\tau)$ have density $f(\tau) = (\tau \times 5\log 10)^{-1}\mathbb{1}_{(0.1\le\tau\le 10^4)}$ which is equivalent to putting the Unif$(-1, 4)$ prior on $\log_{10}\tau$. In our software we use a discrete approximation of this prior induced by drawing $\log_{10}\tau$ uniformly from $\{-1, -0.5, 0, 0.5, 1, \ldots, 4\}$.

Model behavior. In our formulation, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ with a superscript "$*$" are cluster-specific parameters that govern the centroid and the within-cluster variance of each cluster. $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ with a "$\sim$" on top are parameters that determine the centroids and the variability of the shared base distribution. For $i = 1, \ldots, n$, recall that $c_i \in \mathbb{N}^+$ is the cluster label for the ith sample. Moreover, let $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$, $\boldsymbol{c}^*$ the set of distinct values in $\boldsymbol{c}$, and $k^* = |\boldsymbol{c}^*|$ the number of distinct clusters. We note that the actual values of c_i bear no significance and thus assume that the c_i's take integer values between 1 and $|\boldsymbol{c}^*|$. At each node $A \in \mathcal{I}$, $\gamma(A)$ serves as a selector: $\tilde{\theta}(A)$ and $\tilde{\tau}(A)$ become relevant only if $\gamma(A) = 0$. If $\gamma(A) = 1$, they are masked and not used by the model. We note that this masking happens at the level of the base distribution of the Dirichlet process mixture model. If $\gamma(A) = 0$, the base distribution G_0^A is a point mass. Thus, $(\theta_i'(A), \tau_i'(A))$ must be the same, although $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s may not be the same. In a special case, when $\gamma(A) = 0$ for all $A \in \mathcal{I}$, the entire base distribution is a point mass, and $(\boldsymbol{\theta}_i', \boldsymbol{\tau}_i')$'s are all the same. In this case the cluster labels $\boldsymbol{c}$ are only nominal—the samples are from a single cluster although it is possible that $|\boldsymbol{c}^*| > 1$. Similarly, $\gamma(A)$ as an OTU selector is also nominal—A is not necessarily relevant to clustering even if $\gamma(A) = 1$. In real applications, what we care about are not these "nominal" parameters $\boldsymbol{c}$ and $\boldsymbol{\gamma}$ per se, but their "actual" counterparts. Specifically, let $g_i \in \mathbb{N}^+$ be the "actual" cluster label of sample i, and let $s(A) \in \{0, 1\}$ be the "actual" indicator of whether A is relevant to clustering. Moreover, let $\boldsymbol{g} = (g_1, \ldots, g_n)$ and $\boldsymbol{s} = \{s(A) : A \in \mathcal{I}\}$. We have

$$\boldsymbol{g} = \begin{cases} \boldsymbol{c}, & \text{if } \boldsymbol{\gamma} \neq \mathbf{0}_{M-1} \text{ and } \boldsymbol{c} \neq \mathbf{1}_n, \\ \mathbf{1}_n, & \text{if } \boldsymbol{\gamma} = \mathbf{0}_{M-1} \text{ or } \boldsymbol{c} = \mathbf{1}_n, \end{cases} \qquad \boldsymbol{s} = \begin{cases} \boldsymbol{\gamma}, & \text{if } \boldsymbol{c} \neq \mathbf{1}_n, \text{ and } \boldsymbol{\gamma} \neq \mathbf{0}_{M-1}, \\ \mathbf{0}_{M-1}, & \text{if } \boldsymbol{c} = \mathbf{1}_n \text{ or } \boldsymbol{\gamma} = \mathbf{0}_{M-1}. \end{cases} \tag{17}$$

Unlike $\boldsymbol{c}$ and $\boldsymbol{\gamma}$, $\boldsymbol{g}$ and $\boldsymbol{s}$ are directly interpretable. For example, $A \in \mathcal{I}$ is relevant to clustering if and only if $s(A) = 1$. In microbiome applications it is typically expected that the samples have a latent clustering pattern. Therefore, it is common that $\boldsymbol{g} = \boldsymbol{c}$ and $\boldsymbol{s} = \boldsymbol{\gamma}$.

2.4. *Inference strategy.* Under DTMM we are interested in inferring the nominal cluster labels $\boldsymbol{c}$ and the nominal activation indicator $\boldsymbol{\gamma}$ from which the actual cluster labels $\boldsymbol{g}$ and the actual activation indicators $\boldsymbol{s}$ can be obtained. Let $\boldsymbol{y}_{-i}$ denote all observations other than $\boldsymbol{y}_i$. Bayesian inference for DTMM can be achieved by constructing a Markov chain that converges to the joint posterior of $(\boldsymbol{c}, \boldsymbol{\gamma})$. Techniques for Dirichlet process mixture models, such as those described in Neal (2000) or Ishwaran and James (2001), can be applied here.

For $c \in \boldsymbol{c}^*$, let $\boldsymbol{\psi}_c^* = (\boldsymbol{\theta}_c^*, \boldsymbol{\tau}_c^*)$ be the parameters that define the cluster indicated by c (we also let $\boldsymbol{\psi}_c^*(A) = (\theta_c^*(A), \tau_c^*(A))$ for $A \in \mathcal{I}$). Similarly, let $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\tau}})$ be the shared parameters at the coupled nodes and $\tilde{\boldsymbol{\psi}}(A) = (\tilde{\theta}(A), \tilde{\tau}(A))$ for $A \in \mathcal{I}$. The set of unknown parameters in DTMM is $\{\{\boldsymbol{\theta}_i, c_i\}_{i=1}^n, \{\boldsymbol{\psi}_c^*\}_{c=1}^{k^*}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\psi}}, \beta, \lambda\}$. In this work we construct a collapsed Gibbs sampler that iteratively samples from the joint posterior of $(\boldsymbol{c}, \boldsymbol{\gamma}, \beta, \lambda)$. The key to our inference strategy is to compute the marginal likelihoods of samples from a given cluster, integrating out both the sample-specific parameter $\boldsymbol{\theta}_i$ and the cluster-specific parameter $\boldsymbol{\psi}_c^*$. This can be achieved numerically due to two facts. First, the beta-binomial conjugacy makes it easy to integrate out $\boldsymbol{\theta}_i$. Second, the tree-based decomposition of the Dirichlet distribution and the multinomial likelihood provides a divide-and-conquer strategy to marginalize out the high-dimensional cluster-specific parameters $\boldsymbol{\psi}_c^*$ through performing a series of low-dimensional integrals at the internal nodes of the tree.

Specifically, for any $c \in \boldsymbol{c}^*$, let $\boldsymbol{Y}_c^I = \{\boldsymbol{y}_i : c_i = c, i \in I\}$ be a set of samples in cluster c, where $I \subset [n] := \{1, \ldots, n\}$. We also let $\boldsymbol{Y}_c = \boldsymbol{Y}_c^{[n]}$ be the set of all samples in cluster c and $\boldsymbol{Y}_c^{-i} = \boldsymbol{Y}_c^{[n]\setminus\{i\}}$ be the set of samples in cluster c, excluding sample i. For $A \in \mathcal{I}$, let $\mathcal{L}^A(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A), \tilde{\boldsymbol{\psi}}(A))$ be the marginal likelihood of $\boldsymbol{Y}_c^I$ at node A by marginalizing out the sample-specific parameters. The beta-binomial conjugacy yields

$$\begin{aligned}
&\mathcal{L}^A\big(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A), \tilde{\boldsymbol{\psi}}(A)\big) \\
&\quad = \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \frac{B(\theta_c^*(A)\tau_c^*(A) + y_i(A_l), (1-\theta_c^*(A))\tau_c^*(A) + y_i(A_r))}{B(\theta_c^*(A)\tau_c^*(A), (1-\theta_c^*(A))\tau_c^*(A))}.
\end{aligned} \tag{18}$$

We then further integrate out $\boldsymbol{\psi}_c^*(A)$ to obtain the marginal likelihood of $\boldsymbol{Y}_c^I$ at node A, given only the activation indicators and the base parameters,

$$\begin{aligned}
\mathcal{L}_1^A(\boldsymbol{Y}_c^I) &:= \iint \mathcal{L}^A\big(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)\big)\, d\Pi\big(\boldsymbol{\psi}_c^*(A) \mid \gamma(A) = 1, \tilde{\boldsymbol{\psi}}(A)\big) \\
&= \iint \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \\
&\qquad \times \frac{B(\theta(A)\tau(A) + y_i(A_l), (1-\theta(A))\tau(A) + y_i(A_r))}{B(\theta(A)\tau(A), (1-\theta(A))\tau(A))} \\
&\qquad \times \frac{\theta(A)^{\theta_0(A)\nu_0(A)-1}(1-\theta(A))^{(1-\theta_0(A))\nu_0(A)-1}}{B(\theta_0(A)\nu_0(A), (1-\theta_0(A))\nu_0(A))}\, d\theta(A)\, dF^A(\tau),
\end{aligned} \tag{19}$$

$$\begin{aligned}
\mathcal{L}_0^A(\boldsymbol{Y}_c^I \mid \tilde{\boldsymbol{\psi}}(A)) &:= \iint \mathcal{L}^A\big(\boldsymbol{Y}_c^I \mid \boldsymbol{\psi}_c^*(A), \gamma(A) = 0, \\
&\qquad \tilde{\boldsymbol{\psi}}(A)\big)\, d\Pi\big(\boldsymbol{\psi}_c^*(A) \mid \gamma(A) = 0, \tilde{\boldsymbol{\psi}}(A)\big) \\
&= \prod_{\{i \in I : c_i = c\}} \binom{y_i(A)}{y_i(A_l)} \\
&\qquad \times \frac{B(\tilde{\theta}(A)\tilde{\tau}(A) + y_i(A_l), (1-\tilde{\theta}(A))\tilde{\tau}(A) + y_i(A_r))}{B(\tilde{\theta}(A)\tilde{\tau}(A), (1-\tilde{\theta}(A))\tilde{\tau}(A))}.
\end{aligned} \tag{20}$$

Integrals in (19) are two-dimensional integrals that are easy to evaluate numerically. In comparison, to perform a fully Bayesian inference for DMM, the high-dimensional cluster centroids $\boldsymbol{\alpha}_k$'s in (2) have to either be integrated out directly or be sampled in the MCMC procedure. With these marginal likelihoods we can construct our Gibbs sampler for posterior inference. Details on deriving and implementing the Gibbs sampler are given in Section 1.1 and 1.2 of the Supplementary Material (Mao and Ma (2022)).

After running the chain for T iterations, we discard the first B samples as burn-in and obtain $(T - B)$ posterior samples, denoted as $[\{\boldsymbol{c}^{(B+1)}, \boldsymbol{\gamma}^{(B+1)}, \beta^{(B+1)}, \lambda^{(B+1)}\}, \ldots, \{\boldsymbol{c}^{(T)}, \boldsymbol{\gamma}^{(T)}, \beta^{(T)}, \lambda^{(T)}\}]$. Based on these posterior samples, we can compute the posterior samples for $\boldsymbol{g}$ and $\boldsymbol{s}$, based on (17). We denote these posterior samples as $[\{\boldsymbol{g}^{(B+1)}, \boldsymbol{s}^{(B+1)}\}, \ldots, \{\boldsymbol{g}^{(T)}, \boldsymbol{s}^{(T)}\}]$.

For each sample $\boldsymbol{g}^{(t)}$, let $\Gamma^{(t)}$ be the corresponding $n \times n$ association matrix whose (i_1, i_2) element is 1, if $g_{i_1}^{(t)} = g_{i_2}^{(t)}$, and 0 otherwise. Elementwise average of $\Gamma^{(B+1)}, \ldots, \Gamma^{(T)}$ provides an estimation $\hat{\Pi}$ of the pairwise clustering probability matrix Π whose (i_1, i_2) element is $\Pr(\boldsymbol{y}_{i_1}$ and $\boldsymbol{y}_{i_2}$ in the same cluster). To yield a representative clustering, we can report the least-squares model-based clustering (Dahl (2006)), defined as

$$\boldsymbol{C}_{\mathrm{LS}} = \underset{\{\boldsymbol{g}^{(t)}: B<t\le T\}}{\arg\min} \sum_{1\le i_1\le n} \sum_{1\le i_2\le n} \bigl(\Gamma_{i_1 i_2}^{(t)} - \hat{\Pi}_{i_1 i_2}\bigr)^2. \tag{21}$$

$\boldsymbol{C}_{\mathrm{LS}}$ has the advantage that it incorporates information from all posterior samples while output one of the observed clustering in the Markov chain (Dahl (2006)). Other representative clusterings, such as the MAP clustering or the clustering given by the last iteration, can also be used.

Given any representative clustering and the corresponding activation indicators, we can portray the cluster centroids by computing the posterior means of the cluster-specific parameters. Details are provided in Section 1.3 of the Supplementary Material. We also note that the DTMM framework can also be used in the supervised setting to achieve sample classification, based on a training microbiome dataset. Details of classification under the DTMM framework can be found in Section 1.4 of the online Supplementary Material.

3. Numerical experiments.

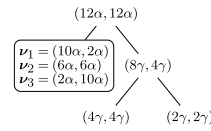
3.1. *Simulation studies*. We first carry out a series of simulation studies to evaluate the performance of DTMM and compare it to several other methods for clustering microbiome count data, namely, the Dirichlet multinomial mixtures (DMM) (Holmes, Harris and Quince (2012)), the k-means algorithm (K-ms) (Lloyd (1982)), the partitioning around medoids algorithm (PAM) (Kaufman and Rousseeuw (2009)), hierarchical clustering (Hclust) (Kaufman and Rousseeuw (2009)) and spectral clustering (Spec) (Ng, Jordan and Weiss (2002)).

3.1.1. *Simulation setup*. In the numerical examples, we simulate datasets with n samples and six OTUs. In each dataset the n samples are denoted as $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i6})$, $i = 1, \ldots, n$, which are generated from the following model:

$$\boldsymbol{y}_i \mid N_i, \boldsymbol{p}_i \overset{\text{ind}}{\sim} \mathrm{Mult}(N_i, \boldsymbol{p}_i) \quad \text{and} \quad \boldsymbol{p}_i \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k \cdot H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k), \tag{22}$$

where the mixture kernel $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k)$ is a distribution on the 5-simplex with parameter $\boldsymbol{\beta}_k$, $N_i \overset{\text{iid}}{\sim} \text{Neg-Binom}(m, s)$. We take the tree in Figure 1 as the "phylogenetic tree" over the six OTUs and consider five different simulation scenarios by choosing different mixture kernels $H_k(\boldsymbol{p}_i \mid \boldsymbol{\beta}_k)$ in (22). In each scenario we let $n = 90$ or 180, $K = 3$ and $(\pi_1, \pi_2, \pi_3) =$

TABLE 2
Mixture kernels for generating the simulated datasets

	Kernel	Signal Level	Signal Parameter	$\boldsymbol{\beta}_k$	
I	DT	W	$\alpha = 1$		$\gamma = 0.1$
		M	$\alpha = 3$		
		S	$\alpha = 6$		
II	Dir	W	$\alpha_0 = 1$	$\boldsymbol{\alpha}_1 = (2, 2, 5, 2, 3, 1) \cdot \alpha_0$	
		M	$\alpha_0 = 3$	$\boldsymbol{\alpha}_2 = (2, 4, 3, 2, 1, 3) \cdot \alpha_0$	
		S	$\alpha_0 = 6$	$\boldsymbol{\alpha}_3 = (2, 6, 1, 2, 2, 2) \cdot \alpha_0$	
III	LN	W	$\alpha = 3$	$q_k = \mathrm{DT}_{\mathcal{T}_6}(\boldsymbol{\nu}_k; \alpha; 0.5)$	$\boldsymbol{\mu}_k = \mathbb{E}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right]$
		M	$\alpha = 6$	$\boldsymbol{\nu}_1 = (10\alpha, 2\alpha)$	$\boldsymbol{\Sigma}_k = \mathbb{V}_{q_k}\left[\log\left(\frac{\boldsymbol{x}_{-6}}{x_6}\right)\right]$
		S	$\alpha = 9$	$\boldsymbol{\nu}_2 = (6\alpha, 6\alpha)$	
				$\boldsymbol{\nu}_3 = (2\alpha, 10\alpha)$	
IV	LN	W	$a = 5, b = 3$	$\boldsymbol{\mu}_1 = (3, 1, a, b, 0)$	$\Sigma_{1,2,3} = \begin{pmatrix} 0.05 & & & & \\ & 0.05 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix}$
		M	$a = 2, b = 2$	$\boldsymbol{\mu}_2 = (2.43, 2.43, a, b, 0)$,	
		S	$a = 1, b = 1$	$\boldsymbol{\mu}_3 = (1, 3, a, b, 0)$	
V	LN	W	$c = 6, d = 6$	$\boldsymbol{\mu}_1 = (c, d, 3.5, 3, 2.5)$	$\Sigma_{1,2,3} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 0.05 & & \\ & & & 0.05 & \\ & & & & 0.05 \end{pmatrix}$
		M	$c = 3, d = 3$	$\boldsymbol{\mu}_2 = (c, d, 2.5, 3.5, 3)$,	
		S	$c = 1, d = 1$	$\boldsymbol{\mu}_3 = (c, d, 3, 2.5, 3.5)$	

$(\frac{4}{9}, \frac{3}{9}, \frac{2}{9})$. Parameters for the negative-binomial distribution are chosen as $m = 15{,}000$, $s = 20$ such that the generated total counts has mean 15,000 and standard deviation 3346, with 95% of them fall into the range (9158, 22258). In the five simulation scenarios the mixture kernels are chosen as shown in Table 2. Details for the simulation setups can be found in Section 2.1 of the Supplementary Material.

In each scenario a "null" case is also considered by setting $K = 1$ in the case with the medium signal level. For each (kernel, signal level) combination we conduct 100 rounds of simulations. For each simulated dataset with $K = 3$, we calculate the following R^2 as a measure of the strength of the signal (Anderson (2001)):

$$R^2 = \frac{\mathrm{SSW}}{\mathrm{SST}} = \frac{\sum_{k=1}^{3}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{\mathrm{BC}}(\boldsymbol{y}_i, \boldsymbol{y}_j)^2 \epsilon_{ij}^k / n_k}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{\mathrm{BC}}(\boldsymbol{y}_i, \boldsymbol{y}_j)^2 / n},$$

where $d_{\mathrm{BC}}(\cdot, \cdot)$ is the Bray–Curtis dissimilarity, n_k the number of samples in cluster k, $\epsilon_{ij}^k = 1$ if the samples i and j are both in cluster k and 0 otherwise. For example, the average R^2s of the 100 simulated datasets in each experiment are reported in Table 3 for $n = 90$.

In each simulation round we run the Gibbs sampler for DTMM for 2000 iterations and discard the first half of the chain as burn-in. The priors and hyperparameters for DTMM are set to the recommended choice in Section 2.3. The initial values for the clustering labels in the Markov chain are set to the labels of running the k-means algorithm with $k = 5$. For DTMM we output $\boldsymbol{C}_{\mathrm{LS}}$ as a representative clustering. For PAM and Hclust we use the Bray–Curtis dissimilarity on the relative abundance as the underlying distance measure between samples. For all competitors other than DMM, the number of clusters is required as a tuning parameter, we set this parameter to the true value 3 when running these methods.

TABLE 3
RMSE of the Jaccard index (small sample size). Cells with the lowest RMSE in each row are highlighted

		$n = 90$							
		Signal		Method					
Expt		Level	R^2	DTMM	DMM	K-ms	PAM	Hclust	Spec
I	DT	–	–	**0.43**	0.51	–	–	–	–
		W	0.30	**0.56**	0.64	0.67	0.71	0.65	0.71
		M	0.35	**0.33**	0.65	0.69	0.69	0.64	0.71
		S	0.37	**0.17**	0.65	0.69	0.71	0.65	0.70
II	Dir	–	–	0.35	**0.00**	–	–	–	–
		W	0.35	**0.53**	**0.53**	0.55	0.59	0.58	0.57
		M	0.52	**0.18**	0.30	0.37	0.33	0.37	0.33
		S	0.60	**0.04**	0.09	0.32	0.19	0.38	0.22
III	LN-A	–	–	0.46	**0.06**	–	–	–	–
		W	0.37	**0.49**	0.64	0.53	0.53	0.54	0.54
		M	0.38	**0.23**	0.64	0.50	0.46	0.55	0.47
		S	0.39	**0.10**	0.64	0.48	0.44	0.53	0.46
IV	LN-S	–	–	0.60	**0.54**	–	–	–	–
		W	0.10	**0.35**	0.72	0.77	0.78	0.73	0.74
		M	0.41	**0.21**	0.54	0.59	0.54	0.60	0.53
		S	0.60	**0.17**	0.37	0.36	0.24	0.41	0.27
V	LN-M	–	–	**0.41**	0.61	–	–	–	–
		W	0.04	**0.20**	0.78	0.78	0.79	0.78	0.76
		M	0.23	**0.14**	0.65	0.76	0.70	0.74	0.68
		S	0.53	**0.17**	0.49	0.22	0.20	0.39	0.22

3.1.2. *Analyses*. To compare the performance of different methods, we compute the Jaccard index (Jaccard (1912)) between the clusters obtained by each method and the true clustering. For a specific clustering $\boldsymbol{c}$ and the true clustering $\boldsymbol{c}_0$, the Jaccard index between $\boldsymbol{c}$ and $\boldsymbol{c}_0$ is defined as $J(\boldsymbol{c}, \boldsymbol{c}_0) = \mathcal{N}_{\boldsymbol{c}\cap\boldsymbol{c}_0}/\mathcal{N}_{\boldsymbol{c}\cup\boldsymbol{c}_0}$, where $\mathcal{N}_{\boldsymbol{c}\cap\boldsymbol{c}_0}$ is the number of pairs of samples that are in the same cluster under both $\boldsymbol{c}$ and $\boldsymbol{c}_0$, $\mathcal{N}_{\boldsymbol{c}\cup\boldsymbol{c}_0}$ the number of pairs of samples that are in the same cluster under at least one of $\boldsymbol{c}$ and $\boldsymbol{c}_0$. When $\boldsymbol{c}$ gives the same clustering as $\boldsymbol{c}_0$, $J(\boldsymbol{c}, \boldsymbol{c}_0) = 1$. In each simulation scenario we compare the root mean squared error of each method m: $\mathrm{RMSE}^{(m)} = \sqrt{\sum_{r=1}^{100}[J(\boldsymbol{c}_r^{(m)}, \boldsymbol{c}_0) - 1]^2/100}$, where $\boldsymbol{c}_r^{(m)}$ is the clustering obtained by method m in simulation round r. As some references, let $\boldsymbol{c}_0 = (1 \cdot \mathbf{1}_{40}^\top, 2 \cdot \mathbf{1}_{30}^\top, 3 \cdot \mathbf{1}_{20}^\top)$, $\boldsymbol{c}_1 = (1 \cdot \mathbf{1}_{90}^\top)$, $\boldsymbol{c}_2 = (1 \cdot \mathbf{1}_{30}^\top, 2 \cdot \mathbf{1}_{30}, 3 \cdot \mathbf{1}_{30})$ and $\boldsymbol{c}_3 = (1 \cdot \mathbf{1}_{40}^\top, 2 \cdot \mathbf{1}_{50})$, where $\mathbf{1}_n$ is the n-dimensional vector with all element equal to 1. We have $\sqrt{[J(\boldsymbol{c}_1, \boldsymbol{c}_0) - 1]^2} = 0.65$, $\sqrt{[J(\boldsymbol{c}_2, \boldsymbol{c}_0) - 1]^2} = 0.50$ and $\sqrt{[J(\boldsymbol{c}_3, \boldsymbol{c}_0) - 1]^2} = 0.30$. The RMSE of DTMM and the competitors under all simulation scenarios when $n = 90$ is shown in Table 3. The RMSE table for $n = 180$ as well as boxplots of the Jaccard index reported by each method in different simulation scenarios can be found in Table S1, Figure S3 and Figure S4 in Section 2.3 of the Supplementary Material.

When $K = 3$, DTMM is always one of the top two methods under comparison. When it is not the best method, its performance is close to the best. Without utilizing the information provided by the phylogenetic tree, all competitors of DTMM suffer when the signal is weak or medium. Moreover, these competitors rely on global distance measures between samples and treat the six OTUs equivalently. As a result, in scenarios like I and IV where the signal is local to a single internal node of the phylogenetic tree, these methods have poor performance. Even in scenario V, where half of the OTUs are relevant for clustering, these methods still

suffer unless the signal is very strong. In scenario II, where the signal is global, all methods perform reasonably well. In this scenario, DTMM can outperform DMM when $n = 90$ even though the latter is the true model. This is because DMM relies on a Laplace approximation to a six dimensional integral when computing the marginal likelihoods to choose the number of clusters. When the sample size is small, DMM tends to choose less than three clusters due to the poor approximation. When $n = 180$, DMM is more likely to choose the right number of clusters, even with the inaccurate approximation. Thus, the performance of DMM improves significantly with more samples. Our experience suggests that DMM tends to underestimate the number of clusters in most cases. For example, in scenarios I and III, DMM usually puts all samples in a same cluster when $n = 90$.

In our simulation settings there are two factors that determine the effect of the increase of sample size on the performance of the two model-based clustering methods. On the one hand, since more samples are available per cluster, the models have a better chance to capture the cluster centroids well once they identify the correct number of clusters. On the other hand, more samples make it harder to get the number of clusters right. These two fighting forces together determine the overall performance shift of the two model-based methods, yet which force prevails is unclear. For DTMM, when the model is misspecified (as in scenarios III, IV and V), the model tends to identify too many small clusters. For the distance-based clustering methods, the second factor plays no role since we assume that the number of clusters is known. In general, our observations suggest that these methods benefit a little from more samples when the signal is strong. Among the distance-based methods, PAM and Spec have a better overall performance. We thus recommend using these two methods to help choose the initial values of DTMM.

We next zoom in to an example to further study the properties of DTMM. In this example we consider a specific simulation round in scenario IV with the medium noise level ($n = 90$). Figure 6 shows the two-dimensional NMDS plot of the samples colored by the clustering obtained by each method. In this example the clustering is roughly determined only by the first NMDS axis. With the node selection module, DTMM is capable of picking the relevant dimensions and clustering efficiently. As for a representative clustering, DTMM finds four clusters, with one falsely identified cluster containing only two samples. This is consistent

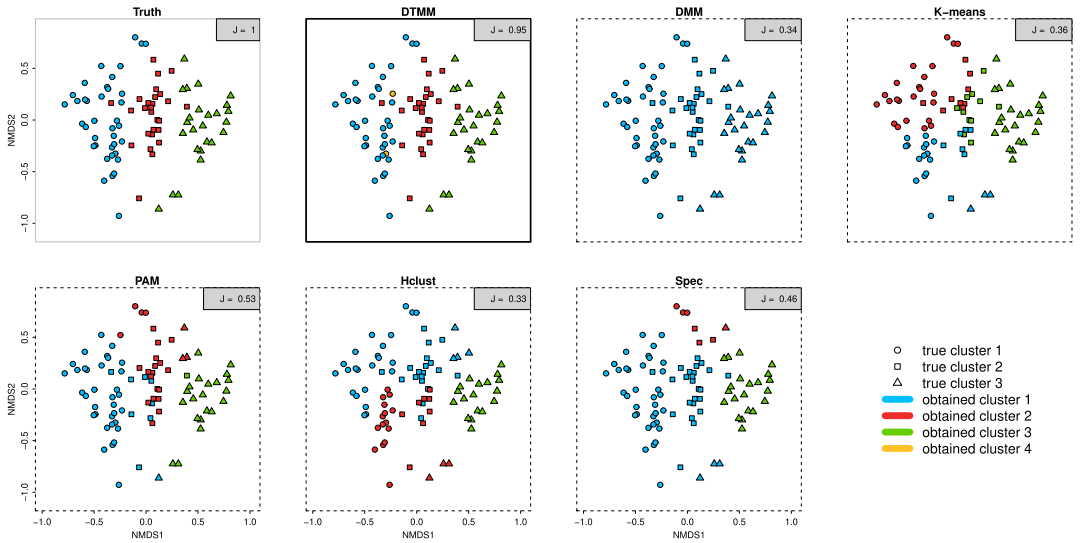


FIG. 6. *2D NMDS plot of samples in a simulation round in scenario IV* ($n = 90$, *medium noise level*). *In each subplot, the true clustering is indicated by the shape of the points while the clustering obtained is indicated by the color.*

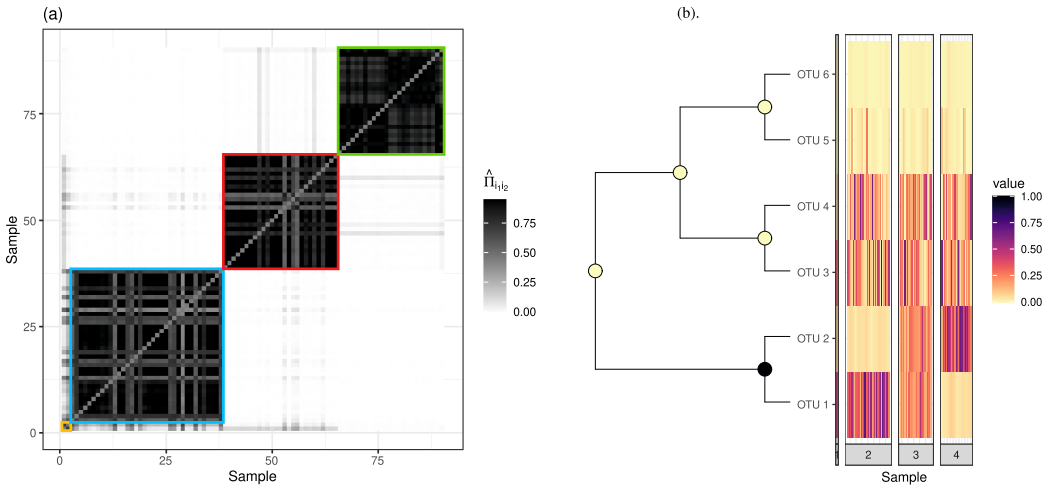


FIG. 7. *Illustrations for an example from simulation scenario IV.* (*a*): *Probability of two samples being clustered together by DTMM, based on* 1000 *post-burnin MCMC samples. The samples are ordered by their cluster labels from DTMM. The clusters identified by DTMM are highlighted by squares colored as in Figure 6.* (*b*): *An illustration of the node selection property of DTMM. The nodes are colored by their estimated posterior node activation probabilities. The heatmap plots the relative abundance of the samples grouped by their cluster labels from DTMM.*

with the well-known fact that inference based on Dirichlet process mixture models can identify small clusters that do not reflect the true data-generating process (Miller and Harrison (2013)). One feature that differs DTMM from its competitors is that it not only outputs a representative clustering but also a whole MCMC trajectory that allows natural uncertainty quantifications. Figure 7(a) shows the probability of two samples being clustered together by DTMM. Clearly, three stable clusters are identified. Although a point estimate from DTMM falsely puts the first two samples in a separate cluster, the uncertainty is large. Figure 7(b) shows the relative abundance of the samples as well as the estimated posterior node activation probabilities. In this example, DTMM is able to uncover the internal nodes that are relevant for clustering. We also consider an example from simulation scenario V. Illustrations similar to Figure 6 and Figure 7 can be found in Figure S5 and S6 in Section 2.3 of the Supplementary Material.

3.2. *Validation.* Validating the results of unsupervised learning is often challenging. In microbiome clustering analyses, the best practice is to check the resulting clusters with scientists to gain biological insights on a case-by-case basis. Instead of trying to provide a general solution of how to justify the clusters found by DTMM, we provide an example to show that DTMM can identify biologically meaningful clusters in real microbiome applications.

Specifically, we reanalyze the data in Dethlefsen and Relman (2011) which studies the responses of stable gut microbiota to antibiotic disturbance. In this study the distal gut microbiome of three patients (patients D, E and F) were monitored over 10 months, including two five-day antibiotic treatment courses separated by a five-month interim period. Fifty-two to fifty-six samples were collected for each patient in the experiment. Samples of patients D and F are shown in Figure 8 and Figure 9 which also illustrate the design of the study. In our analysis we aggregate the OTU counts to the genus level which gives 59 OTUs in total.

As in Dethlefsen and Relman (2011), we analyze the samples from the three patients separately. For each patient we ignore the time information of when the samples were taken and run DTMM on these samples for 2500 iterations. The first half of the chain was discarded as burn-in. The clustering results for patient D are shown in Figure 8 (the x-axis labels in these plots are colored by the cluster labels in $\boldsymbol{C}_{\mathrm{LS}}$ of the samples they represent). For this patient,

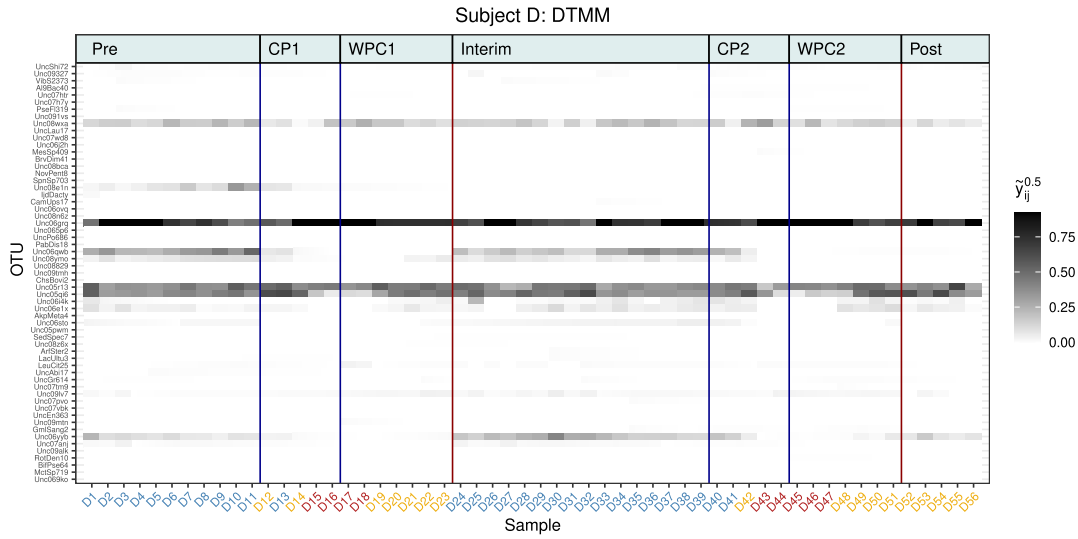


FIG. 8. *The heatmap of the microbiome samples of patient D (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DTMM. The blue vertical lines mark the two antibiotic treatment courses. "CP" denotes the antibiotic treatment (ciprofloxacin); "WPC" is the week posttreatment; "Pre" and "Post" denote the pretreatment and posttreatment periods.*

DTMM identifies three clusters which can be interpreted as the *stable*, *sterile* and *recover* stages of the microbiota. Based on the clustering results, the gut microbiota of patient D was stable before the treatment. It was able to recover to some stable states from antibiotic treatment within a week after the treatment was finished. However, although the microbiota was able to fully recover to the pretreatment state after the first antibiotic treatment course, it never made a full recovery to the original state after the second (repeated) antibiotic treatment.

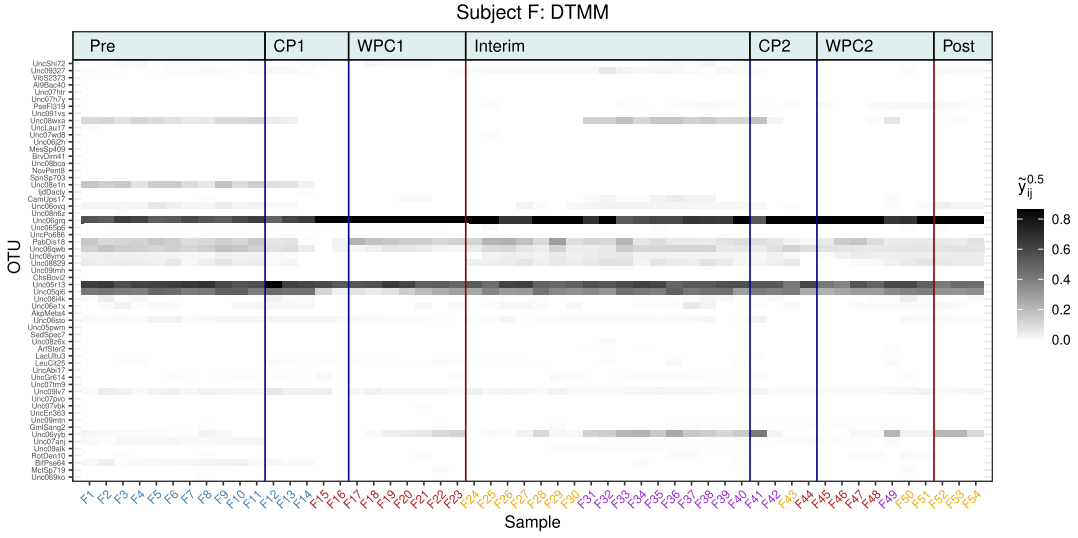


FIG. 9. *The heatmap of the microbiome samples of patient F (after the square-root transform). Each column represents a specific sample. The columns are ordered by the times the samples were collected. The colors of the x-axis labels represent the clustering labels of the samples returned by DTMM. The legends are defined the same way as in Figure 8.*

Similarly, the results for patient F are provided in Figure 8. For patient F, DTMM identifies four clusters corresponding to the *sterile*, *recover* and two different *stable* stages, respectively. Like patient D, the gut microbiota of patient F was stable before the treatment and was able to recover from the treatments. Unlike patient D, it did not recover to the pretreatment state, even after the first treatment course. Moreover, it took longer for patient F to recover than patient D. We note that these findings are all consistent to the findings in Dethlefsen and Relman (2011), where the time and design information was used to get these results.

As a comparison, the clustering results for these two patients under DMM are shown in Figure S7 and Figure S8 of the Supplementary Material 2.3. For both patients, DMM returns two clusters that roughly represent the *stable* and *unstable* stages of the microbiota. In this example, DTMM is able to discover more interesting latent structures among samples than DMM. It is worth noting that in each analysis, microbiome samples were collected from the same patient. Thus, the level of cross-sample variations in this study is much smaller than microbiome studies with multiple subjects. In those cases, we expect DTMM to benefit more from its improved flexibility over DMM and discover even more interpretable structures than the latter.

4. Case studies. The American Gut Project (McDonald, Birmingham and Knight (2015), McDonald et al. (2018)) aims at building an open-source and open-access reference microbiome dataset for general scientific use, based on 16S rRNA sequencing and the QIIME pipeline (Caporaso et al. (2010)). It collects mouth, skin and feces samples over a large variety of U.S. participants on a voluntary basis. The participants send their microbiome samples to UC San Diego for sequencing and complete a questionnaire that covers their dietary habits, lifestyle and health history.

We apply DTMM to the July 2016 version of the fecal data from the AGP to construct enterotypes for two groups of samples: first, we consider participants who have been diagnosed with inflammatory bowel disease (IBD); second, we consider participants who have been diagnosed with diabetes. The diagnoses are made by a medical professional (a doctor or a physician assistant). The specific version of the AGP dataset contains an OTU table of 27,774 OTUs. We focus on the top 75 OTUs, based on total counts, to reduce noises in the dataset and control for the sequencing errors. The top 75 OTUs, on average, retain two-thirds of the total counts in a sample. We filter the samples by only considering participants with at least 500 counts on the top 75 OTUs. This filtering ends up with 189 samples diagnosed with IBD and 106 samples diagnosed with diabetes.

In the following sections we fit DTMM to each or the two datasets with the priors and hyperparameters set to the recommended choices in Section 2.3. In each analysis we run the Gibbs sampler in Section 2.4 for 5000 iterations and discard the first half of the chain as burn-in. The cluster labels are initiated by running the PAM algorithm with $K = 5$.

Key findings from our analyses of these datasets are summarized as follows: (i) enterotypes of the two disease-diagnosed groups are determined by a large number of OTUs jointly in a sophisticated manner instead of by a few dominant OTUs; (ii) OTUs from genera *Bacteroides, Prevotella* and *Ruminococcus* are typically important in identifying those enterotypes which is consistent to the findings in previous works (Arumugam et al. (2011)); (iii) the number of enterotypes and the OTUs that characterize each enterotype can differ across datasets; and (iv) DMM tends to find larger clusters that are unions of clusters found by DTMM.

4.1. *IBD*. We first consider samples from participants that are diagnosed with IBD. Figure S9 in Section 3 of the Supplementary Material shows the traceplots of some one-dimensional parameters or summaries of the posterior samples. The Markov chain stabilizes and mixes reasonably well after about 750 iterations. Figure S9(a) and Figure S9(b) show

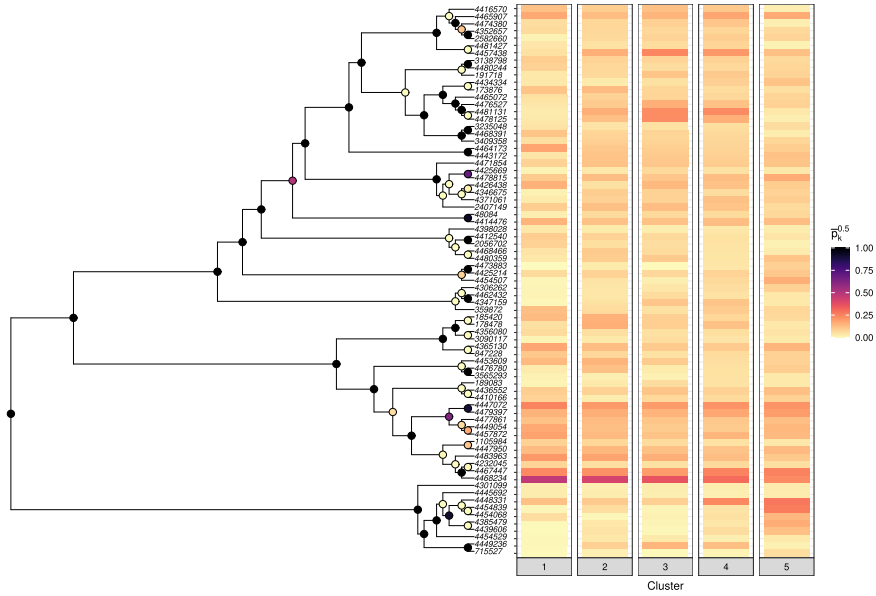


FIG. 10. *Left*: *Estimated posterior means of the activation indicators at each node of the phylogenetic tree. Right*: *The estimated centroids of the five clusters in* $\boldsymbol{C}_{\mathrm{LS}}$ (*after the square-root transform*).

the traceplots of the Dirichlet process precision parameter β and the prior global activation probability λ. The posterior means of these parameters are 0.87 and 0.53, respectively. Traceplot of the sampled number of internal nodes with $\gamma(A) = 1$ is shown in Figure S9(c). On average, 39 out of 74 nodes are marked as relevant to the clustering process, indicating that the clustering process is determined by various OTUs jointly in a complicated way, instead of being dominated by a few OTUs that are abundant in counts. Figure S9(d) shows the cumulative proportion of samples in the largest one, two, three, four and five clusters for each iteration. DTMM tends to assign samples into five clusters.

We find $\boldsymbol{C}_{\mathrm{LS}}$ as defined in (21) which corresponds to $\boldsymbol{c}^{(t_0)}$ with $t_0 = 2717$. $\boldsymbol{C}_{\mathrm{LS}}$ assigns the samples into five clusters with sizes 6, 41, 73, 42 and 27, respectively. The estimated centroids of the five clusters are shown in Figure 10 which also shows the estimated posterior means of the activation indicator $s(A)$ at $A \in \mathcal{I}$. Most internal nodes that are irrelevant to clustering are close to the leaves of the tree. Nodes that are more "global" (have more descendant OTUs) generally contribute to the clustering. This indicates that the clustering process is determined by most OTUs jointly in a complicated manner. Figure 11 (left) shows the estimated pairwise clustering probability matrix $\hat{\Pi}$ with the rows and columns ordered by the labels in $\boldsymbol{C}_{\mathrm{LS}}$. There are noticeable uncertainties in the clustering, especially between clusters 2, 3 and 4. The similarities of these three clusters can also be seen from Figure 11 (right), where we plot the heatmap of the samples (after the square-root transform) grouped by their labels in $\boldsymbol{C}_{\mathrm{LS}}$. Figure 11 (right) also shows that the within-cluster variations among samples are large.

To see which OTUs are more important in determining $\boldsymbol{C}_{\mathrm{LS}}$, we consider the following heuristic measure of OTU importance: for $1 \le j \le M$, let

$$\vartheta_j = \frac{SSB_j}{SSW_j} = \frac{\sum_{c \in \boldsymbol{C}_{\mathrm{LS}}} n_c (\bar{y}_{cj} - \bar{y}_j)^2}{\sum_{c \in \boldsymbol{C}_{\mathrm{LS}}} \sum_{c_i = c} (y_{ij} - \bar{y}_{cj})^2}, \tag{23}$$

where $\bar{y}_j$ is the overall mean of y_{ij}, $\bar{y}_{cj}$ the mean of y_{ij} for samples with $c_i = c$. Table 4 shows the top 10 OTUs in determining $\boldsymbol{C}_{\mathrm{LS}}$ in terms of ϑ_j as well as their compositions in each cluster centroid. Overall, $\boldsymbol{C}_{\mathrm{LS}}$ is jointly determined by multiple OTUs in a complicated way. Note that OTUs that are important for clustering are not necessarily those with abundant

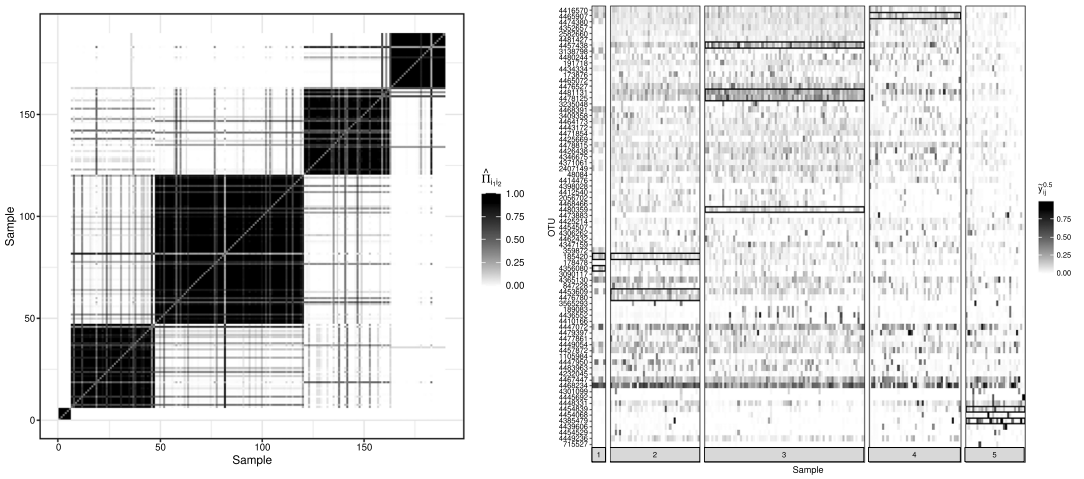


FIG. 11. *Left*: *Estimated pairwise coclustering probabilities*. *Right*: *Heatmap of the samples* (*after the square–root transform*) *grouped by their labels in* $\boldsymbol{C}_{\mathrm{LS}}$. *The black boxes illustrate the characteristic OTUs of each cluster.*

counts. For example, OTU-4468234 and OTU-4447072 (both are *Bacteroides*) are the two OTUs with the most counts in the dataset. However, these two OTUs are prevalent in most samples and thus have limited roles in the clustering.

We next compare the five resulting clusters in more details. Figure 12 shows the boxplot of the Shannon diversity of samples in the five clusters, respectively. Samples from clusters 2 and 3 tend to have more evenly distributed counts across OTUs, compared to those from clusters 1, 4 and 5. Similar to (23) we can define a heuristic measure of OTU importance in characterizing each of the five clusters. Specifically, for $c = 1, \ldots, 5$, let

$$\vartheta_j^c = \frac{SSB_j^c}{SSW_j^c} = \frac{n_c(\bar{y}_{cj} - \bar{y}_j)^2 + n_{-c}(\bar{y}_{-cj} - \bar{y}_j)^2}{\sum_{c_i=c}(y_{ij} - \bar{y}_{cj})^2 + \sum_{c_i \neq c}(y_{ij} - \bar{y}_{-cj})^2}, \tag{24}$$

where n_{-c} is the number of samples that are not in cluster c, $\bar{y}_{-cj}$ the mean of y_{ij} for samples with $c_i \neq c$. (24) is equivalent to merging the four clusters, other than cluster c in (23). The boxes in Figure 11 (right) indicate the top OTUs in terms of ϑ_j^c for each c (only OTUs with $\vartheta_j^c > 0.1$ are shown).

Based on these results, we can characterize each cluster by a few OTUs with the top ϑ_j^c. For example, samples from cluster 2 tend to have more counts from the *Rikenellaceae* family

TABLE 4

Estimated cluster-specific compositions of the top 10 *OTUs in determining* $\boldsymbol{C}_{\mathrm{LS}}$ *in terms of* ϑ_j. *Values of the OTU compositions are shown in the percentage scale*

OTU	Family	Genus	ϑ_j	C1	C2	C3	C4	C5
185420	Bacteroidaceae	Bacteroides	0.43	2.01	2.62	0.98	0.66	0.54
4478125	Ruminococcaceae	Faecalibacterium	0.43	0.22	1.71	5.72	2.71	0.10
4356080	Barnesiellaceae	–	0.33	0.53	0.41	0.37	0.42	0.28
4476780	Rikenellaceae	–	0.32	0.14	1.70	0.15	0.33	1.04
4453609	Rikenellaceae	–	0.26	2.08	2.43	1.15	0.71	0.85
4480359	Ruminococcaceae	–	0.22	0.22	1.02	1.22	0.11	1.48
4465907	Lachnospiraceae	Blautia	0.21	3.32	1.82	2.40	3.47	3.04
4481131	Ruminococcaceae	Faecalibacterium	0.18	0.11	2.92	5.62	6.11	0.22
4457438	Lachnospiraceae	–	0.19	0.36	2.72	6.43	4.67	1.68
4385479	Enterobacteriaceae	Proteus	0.17	0.02	0.21	0.04	0.24	2.91

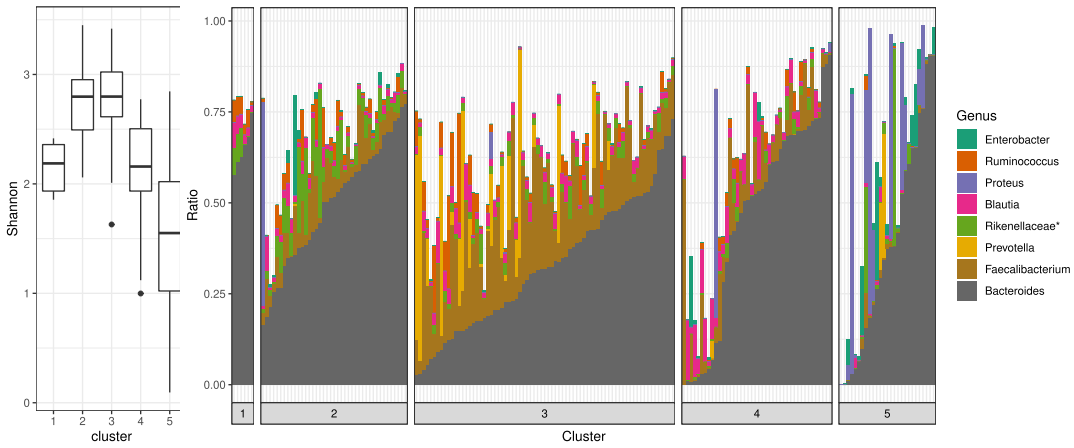


FIG. 12. *Left*: *Boxplot of the Shannon diversity of samples in each cluster. Right*: *Relative abundance of eight genera for each sample. A genus is chosen if its descendant OTUs have large* ϑ_j^c *for some* c. *For the three OTUs with unavailable genera information, their family is shown instead* (*indicated by Rikenellaceae**). *The samples are grouped by their cluster labels in* $\boldsymbol{C}_{\mathrm{LS}}$.

(represented by OTU-4453609 and OTU-4476780). Cluster 3 is characterized by having more abundance in the *Faecalibacterium* (represented by OTU-4478125 and OTU-4481131) and the *Lachnospiraceae* family (represented by OTU-4481127 and OTU-4457438). Arumugam et al. (2011) proposed three enterotypes in human gut microbial communities that are characterized by the variation in the levels of one of the three genera: *Bacteroides, Prevotella* and *Ruminococcus*. Our analysis suggests that enterotypes of the IBD patients are determined by a sophisticated mechanism involving more genera. That said, although OTUs from the *Bacteroides, Prevotella* and *Ruminococcus* genera are not always those with the largest ϑ_j^c, they are playing important roles in identifying the five clusters. For example, OTUs from the *Prevotella* genus have large ϑ_j^3 and are thus crucial in determining cluster 3 while OTUs from the *Ruminococcus* genus have large ϑ_j^1 and ϑ_j^2 and are thus important in identifying cluster 1 and 2. This can also be seen from Figure 12 (right), where relative abundance of 8 genera picked by ϑ_j^c are shown for each sample.

4.2. *Diabetes*. Similar to Section 4.1, we apply DTMM to samples from diabetes patients. Results for this application are provided in Section 3 of the Supplementary Material. For example, counterparts of Figure S9 and Figure 11 are shown in Figure S10 and Figure S11. In this example, DTMM finds three clusters with $\boldsymbol{C}_{\mathrm{LS}}$. Figure S12 shows the estimated centroids of the three clusters as well as the estimated posterior means of the activation indicators at each node of the phylogenetic tree. Figure S13 (right) shows for each sample the relative abundance of six genera selected based on the importance of their descendant OTUs in identifying the three clusters. Compared with the IBD example, enterotypes in this case can be associated with individual OTUs in a simpler manner. For example, samples in cluster 3 tend to have significantly lower abundance in *Faecalibacterium* and *Bacteroides* which are the dominating genera in most samples. Compared with cluster 2, cluster 1 is identified with relatively more counts from OTU-173876 and the *Prevotella* family.

On average, 21 out 75 internal nodes of $\mathcal{T}$ are estimated as relevant to clustering. Compared with the IBD example in Section 4.1, fewer nodes are involved, suggesting that clusters in the diabetes example are determined by fewer OTUs (genera). As shown in Figure S11 (right), a few OTUs play crucial roles in determining multiple clusters. As a comparison, as shown in Figure 11 (right), each cluster in the IBD example is determined by a unique set of

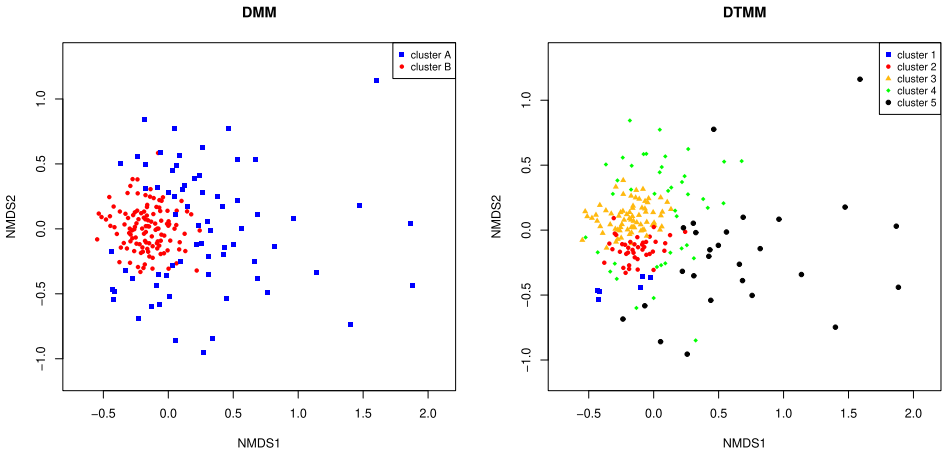


FIG. 13. *Two-dimensional NMDS plots for the AGP IBD dataset. Points are colored and shaped by the clustering given by DMM (left) or DTMM (right).*

OTUs. Since DTMM marks an internal node as relevant if it is relevant in determining *any* cluster, more nodes are selected in the IBD example.

4.3. *DTMM vs. DMM.* We also apply DMM to the two examples in this section and compare it with DTMM. DMM reports two clusters in both examples (see Figure S14 in Section 3 of the Supplementary Material). For the IBD dataset, Figure 13 (left) shows the two-dimensional NMDS plot of the data, colored by the cluster labels reported by DMM. In comparison, Figure 13 (right) shows the same NMDS plot colored by the cluster labels in $\boldsymbol{C}_{\mathrm{LS}}$ reported by DTMM. Similar NMDS plots for the diabetes example are shown in Figure S15 in Section 3 of the Supplementary Material.

For the IBD application, the five clusters reported by DTMM can be seen as refinements of the two clusters reported by DMM. Roughly, cluster B identified by DMM is the union of clusters 2 and 3 from DTMM, while cluster A found by DMM is the union of clusters 1, 4 and 5 from DTMM. As shown in Figure 11 (right) and Figure 12 (right), those subclusters from DTMM are not differentiated by OTUs with dominant counts. Thus, it is very unlikely for DMM to make further splits. Moreover, based on Figure 11 (right), samples within each cluster from DTMM tend to show different levels of heterogeneities across OTUs, making the underlying Dirichlet-multinomial model of DMM unrealistic. For example, counts of OTUs from the *Prevotella* genus tend to show large within cluster variations among samples in cluster 3. To capture this level of variation, DMM has to push the cluster-specific dispersion parameter very large, essentially loose its ability to effectively find those subclusters.

5. Concluding remarks. We have introduced DTMM as a model-based framework for clustering the amplicon sequencing data in microbiome studies. By directly incorporating the phylogenetic tree, DTMM differs from the popular DMM in three directions: first, it offers a more flexible covariance structure among different OTUs; second, it provides a way for selecting a subset of internal nodes in the phylogenetic tree that is relevant for clustering; moreover, it allows simple and efficient algorithms for posterior inference. That said, DTMM does have a higher computational cost than methods such as DMM that only incur simple closed-form conjugate updates, due to the large number of numerical integrals in the form of (19) in the Gibbs sampling. For example, on a single 2.5 GHz Intel Core-i7 desktop core, the validation study takes about six hours to run (on the samples from a patient), and the two case studies take about 48 hours each to run with our current software implementation which

utilizes no parallelization and involves only vanilla grid-based evaluation of the integrals. Such numerical integration can be substantially sped up through approximative strategies, such as Laplace approximation applied on the inner integral on $\theta(A)$ in (19), as proposed in Ma and Soriano (2018), as well as hardware-based parallelization using GPUs. Moreover, the evaluation of the integrals over different nodes on the phylogenetic tree is embarrassingly parallel and thus can be computed simultaneously with multiple cores. We expect future versions of the software for DTMM to incorporate these functionalities that will substantially improve the computational time.

Finally, while the covariance structure offered by DT is richer than that of the Dirichlet distribution, it is still limited compared to the logistic-normal family (LN). In a case with K OTUs, DT models the covariance among OTU counts with $(K-1)$ dispersion parameters in the series of beta distributions while LN uses $K(K-1)/2$ parameters in modeling the covariance matrix. It is interesting to further generalize the covariance structure provided by DTMM without making the inference too complicated. When selecting a subset of internal nodes in the phylogenetic tree that are relevant to clustering, DTMM selects a node if it is relevant in identifying *any* cluster. Intuitively, DTMM first selects a subspace in the node space and performs clustering in that space. An alternative direction worth exploring is to allow the nodes selected to be cluster-dependent such that each cluster can deviate from the "mean" cluster at different internal nodes.

Funding. L. Ma's research is partly supported by NIGMS Grant R01-GM135440 and NSF grant DMS-1749789.

Part of this work was completed when J. Mao was supported by a Duke Forge Graduate Fellowship in health data science.

SUPPLEMENTARY MATERIAL

Supplement A to "Dirichlet-tree multinomial mixtures for clustering microbiome compositions" (DOI: 10.1214/21-AOAS1552SUPPA; .pdf). This supplementary file provides details on the posterior inference and MCMC sampling procedure for DTMM, additional details and results of the numerical examples, and additional results of the case studies.

Supplement B to "Dirichlet-tree multinomial mixtures for clustering microbiome compositions" (DOI: 10.1214/21-AOAS1552SUPPB; .zip). We provide an `R` package (`DTMM`: https://github.com/MaStatLab/DTMM) implementing the proposed method in this paper. We also provide code and a guide to reproduce all simulations and data analyses in Section 3 and Section 4.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

ARUMUGAM, M., RAES, J., PELLETIER, E., LE PASLIER, D., YAMADA, T., MENDE, D. R., FERNANDES, G. R., TAP, J., BRULS, T. et al. (2011). Enterotypes of the human gut microbiome. *Nature* **473** 174.

CALLAHAN, B. J., MCMURDIE, P. J. and HOLMES, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11** 2639–2643. https://doi.org/10.1038/ismej.2017.119

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13** 581.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7** 335.

COSTEA, P. I., HILDEBRAND, F., ARUMUGAM, M., BÄCKHED, F., BLASER, M. J., BUSHMAN, F. D., DE VOS, W. M., EHRLICH, S. D., FRASER, C. M. et al. (2018). Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3** 8–16.

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference Gene Expr. Proteomics* **4** 201–218.

DENNIS, S. Y. III (1991). On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Comm. Statist. Theory Methods* **20** 4069–4081. MR1158563 https://doi.org/10.1080/03610929108830757

DETHLEFSEN, L. and RELMAN, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **108** 4554–4561.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** e30126. https://doi.org/10.1371/journal.pone.0030126

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729 https://doi.org/10.1198/016214501750332758

JACCARD, P. (1912). The distribution of the flora in the Alpine zone. 1. *New Phytol.* **11** 37–50.

KARLSSON, F. H., TREMAROLI, V., NOOKAEW, I., BERGSTRÖM, G., BEHRE, C. J., FAGERBERG, B., NIELSEN, J. and BÄCKHED, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498** 99–103. https://doi.org/10.1038/nature12198

KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding Groups in Data*: *An Introduction to Cluster Analysis* **344**. Wiley, New York.

KNIGHTS, D., KUCZYNSKI, J., CHARLSON, E. S., ZANEVELD, J., MOZER, M. C., COLLMAN, R. G., BUSHMAN, F. D., KNIGHT, R. and KELLEY, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8** 761.

KOREN, O., KNIGHTS, D., GONZALEZ, A., WALDRON, L., SEGATA, N., KNIGHT, R., HUTTENHOWER, C. and LEY, R. E. (2013). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9** e1002863. https://doi.org/10.1371/journal.pcbi.1002863

KOSTIC, A. D., XAVIER, R. J. and GEVERS, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology* **146** 1489–1499.

KUNTZ, T. M. and GILBERT, J. A. (2017). Introducing the microbiome into precision medicine. *Trends Pharmacol. Sci.* **38** 81–91. https://doi.org/10.1016/j.tips.2016.10.001

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7** e52078.

LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489

LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71** 8228–8235.

MA, L. and SORIANO, J. (2018). Analysis of distributional variation through graphical multi-scale beta-binomial models. *J. Comput. Graph. Statist.* **27** 529–541. MR3863755 https://doi.org/10.1080/10618600.2017.1402774

MAO, J. and MA, L. (2022). Supplement to "Dirichlet-tree multinomial mixtures for clustering microbiome compositions." https://doi.org/10.1214/21-AOAS1552SUPPA, https://doi.org/10.1214/21-AOAS1552SUPPB

MAO, J., CHEN, Y. and MA, L. (2020). Bayesian graphical compositional regression for microbiome data. *J. Amer. Statist. Assoc.* **115** 610–624. MR4107661 https://doi.org/10.1080/01621459.2019.1647212

MCDONALD, D., BIRMINGHAM, A. and KNIGHT, R. (2015). Context and the human microbiome. *Microbiome* **3** 52. https://doi.org/10.1186/s40168-015-0117-2

MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C. et al. (2018). American gut: An open platform for citizen science microbiome research. *MSystems* **3** e00031–18.

MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* 199–206.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299 https://doi.org/10.1093/biomet/49.1-2.65

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55.

QUINCE, C., LUNDIN, E. E., ANDREASSON, A. N., GRECO, D., RAFTER, J., TALLEY, N. J., AGREUS, L., ANDERSSON, A. F., ENGSTRAND, L. et al. (2013). The impact of Crohn's disease genes on healthy human gut microbiota: A pilot study. *Gut* **62** 952–954.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

TANG, Y., MA, L. and NICOLAE, D. L. (2018). A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *Ann. Appl. Stat.* **12** 1–26. MR3773384 https://doi.org/10.1214/17-AOAS1086

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457** 480.

WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. MR3713113 https://doi.org/10.1111/biom.12654

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.