A Bayesian Nonparametric Approach for Identifying Differentially Abundant Taxa in Multigroup Microbiome Data with Covariates

Archie Sachdeva
Bristol Myers Squibb, Princeton, NJ
and
Somnath Datta
Department of Biostatistics, University of Florida
and
Subharup Guha
Department of Biostatistics, University of Florida

December 29, 2023

Abstract

Scientific studies in the last two decades have established the central role of the microbiome in disease and health. Differential abundance analysis seeks to identify microbial taxa associated with sample groups defined by a factor such as disease subtype, geographical region, or environmental condition. The results, in turn, help clinical practitioners and researchers diagnose disease and develop treatments more effectively. However, microbiome data analysis is uniquely challenging due to high-dimensionality, sparsity, compositionally, and collinearity. There is a critical need for unified statistical approaches for differential analysis in the presence of covariates. We develop a zero-inflated Bayesian nonparametric (ZIBNP) methodology that meets
these multipronged challenges. The proposed technique flexibly adapts to the unique data characteristics, casts the high proportion of zeros in a censoring framework, and mitigates high-dimensionality and collinearity by utilizing the dimension-reducing property of the semiparametric Chinese restaurant process. Additionally, the ZIBNP approach relates the microbiome sampling depths to inferential precision while accommodating the compositional nature of microbiome data. Through simulation studies and analyses of the CANine Microbiome during Parasitism (CAMP) and Global Gut microbiome datasets, we demonstrate the accuracy of ZIBNP compared to established methods for differential abundance analysis in the presence of covariates.

*Keywords:* Censoring; Chinese restaurant process; Compositional data; Stochastic imputation; MCMC; ZIBNP


## 1 Introduction

Numerous scientific studies have established the influence of the microbiome on health and disease [Wade, 2013, Shreiner et al., 2015, Rautava, 2016]. Giant strides in next-generation sequencing technologies offer a glimpse into the microbiome present in our bodies by facilitating the simultaneous identification and quantification of taxa abundances [Tang et al., 2019], where a taxon represents any profiled set of microbiome features such as operational taxonomic units (OTUs), amplicon sequence variants (ASVs), or metagenomic species. Differential abundance analysis, or the identification of microbial taxa associated with population groups, plays a key role in disease diagnoses and the development of novel treatments [Cappellato et al., 2022]. Informally, a taxon is not differentially abundant if, after adjusting for the covariates, every pair of samples belonging to different groups have similar normalized abundances for that taxon; otherwise, the taxon is differentially abundant. This is especially challenging because microbiome abundance data are high-dimensional, sparse, and compositional, necessitating sophisticated statistical methods [e.g., Weiss et al., 2017]. It is also essential to adjust for individual-specific attributes in differential abundance analyses [Vujkovic-Cvijin et al., 2020]; for example, when the groups are disease subtypes, covariates such as diet and lifestyle that influence microbial composition may be highly associated with the grouping variable.

Microbiome samples are sequenced using one of two processes: amplicon sequencing and random shotgun sequencing. Shotgun sequencing is restricted to highly variable regions of the 16S rRNA gene [Kembel et al., 2012], whereas amplicon sequencing assesses genome-wide genetic variation. Shotgun sequencing has better species-level resolution and ability to detect novel viruses than amplicon sequencing [Bibby, 2013, Relman, 2013], but requires considerably larger volumes [Sharpton, 2014]. We refer the reader to Bharti and Grimm [2019] for a detailed review of sequencing technologies and bioinformatics pipelines for analyzing raw microbiome data [Xia et al., 2018a]. Subsequently, taxonomic profile generation [Anyaso-Samuel et al., 2021, Bharti and Grimm, 2019] maps the microbiome sequences to known reference databases and produces a high-dimensional vector of abundance counts for each sample [Schloss and Westcott, 2011]. Each element of this vector corresponds to a
taxon. The abundance counts of the samples are arranged in a matrix called the *abundance table*; each element represents the measured abundance for a taxon (column) in a sample (row). The number of taxa far exceeds the number of samples, resulting in a “short and wide” abundance table.

The term “abundance counts” is somewhat misleading because high-throughput sequencing technologies do not measure the actual counts. Instead, the counts are measured relative to the total number of reads, commonly referred to as the *library size* or *sampling depth*, related to measurement precision [e.g., Weiss et al., 2017]. After normalization, the sample-specific measurements (i.e., abundance matrix rows) are intrinsically *compositional* and represent random points on a high-dimensional simplex [Aitchison, 2008]. Furthermore, due to technical variations of the sequencing procedure, the library sizes vary from sample to sample [McKnight et al., 2019, Weiss et al., 2017]. Microbiome data are also highly *sparse* and comprise 10% - 70% zeros. Some zeros are caused by individual traits; for instance, low-fiber diets are known to deplete certain gut bacteria [Bailén et al., 2020]. However, most zeros occur due to sequencing process errors or relatively small sampling depths. In the literature, these three types of zeros are respectively referred to as biological, sampling, and technical zeros [Silverman et al., 2020, Kaul et al., 2017, Weiss et al., 2017].

An overview of state-of-the-art, and predominantly frequentist, methods for differential abundance analysis can be found in Wallen [2021] and Datta et al. [2021]. These methods are broadly classified as *compositional* or *count-based* [Liu et al., 2021, Nearing et al., 2022]. Prominent among compositional approaches are ALDEx2 [Fernandes et al., 2014] and ANCOM-II [Kaul et al., 2017], both of which use Wilcoxon rank-sum tests to detect the differential taxa of the groups. ALDEx2 relies on the centered log-ratio (clr) transformation and Dirichlet-multinomial model; ANCOM-II uses the additive log-ratio (alr) transformation. By contrast, count-based approaches utilize the sampling depths to inform inferential precision. Zero-inflated count approaches that attempt to account for sparsity include metagenomeseq [Paulson et al., 2013], which relies on a zero-inflated log-normal model, and Risso et al. [2018] and Xia et al. [2018b], which account for overdispersion using negative binomial models. ANCOMBC [Lin and Peddada, 2020] uses a log-linear regression framework with a bias-corrected random intercept. For two-group comparisons, MaAsLin2 [Mallick
et al., 2021] uses GLMs along with different normalization and distributional options. The methods DESeq2 [Love et al., 2014] and corncob [Martin et al., 2020] respectively utilize negative binomial and beta-binomial distributions and assume marginal distributions for the featurewise counts. However, no statistical method addresses the disparate challenges of microbiome data or performs reliably with different microbiome datasets [Weiss et al., 2017].

We propose a coherent statistical framework called the zero-inflated Bayesian nonparametric (ZIBNP) model that detects differentially abundant taxa of multiple groups of samples while effectively confronting the multipronged challenges of high-dimensionality, compositionality, collinearity, sparsity, and covariate confounding. As its name suggests, the hierarchical model comprises a two-component mixture model for the abundance counts with one mixture component being a point mass at zero and the other component comprising a flexible Bayesian nonparametric distribution whose posterior adapts to the dataset characteristics. Technical and sampling zeros, introduced earlier, are cast in a censoring framework and inferred a posteriori. ZIBNP is a hybrid of compositional and count-based methods because it relates the sample covariates and library sizes to inferential precision while also accommodating the compositional aspect of microbiome data. High-dimensionality and collinearity (specifically, small \( n \), large \( p \) issues) are mitigated by allocating the large number of taxa to fewer latent clusters defined by shared relative abundance patterns using the semiparametric Chinese restaurant process [Müller and Mitra, 2013, Lijoi and Prünster, 2010]. We apply the ZIBNP technique to infer the differentially abundant taxa in two publicly available microbiome datasets: (i) The CAnine Microbiome during Parasitism (CAMP) study [Oliveira et al., 2017] explores the impact of natural parasite infection on the gut microbiome of infected and uninfected domesticated dogs as the two groups; and (ii) Global Gut microbiome data [Yatsunenko et al., 2012] on humans residing in three geographical regions as the groups.

The paper is organized as follows. Section 2 introduces the ZIBNP model. Section 2.1 develops the nonparametric mixture component for mitigating high-dimensionality and collinearity, incorporating covariate effects, and detecting differential taxa. Section 2.2 addresses the challenges of sparsity by fostering a censoring framework for non-biological
zeros. Section 3 outlines the posterior inference procedure. Section 4 uses artificial datasets to demonstrate the high accuracy achieved by ZIBNP relative to existing approaches that adjust for sample covariates. Section 5 analyzes the motivating microbiome datasets using the proposed ZIBNP method. Section 6 concludes with a brief discussion.

2 A Bayesian hierarchical approach

We foster a Bayesian framework capable of inferring the complex relationships between microbial taxa and groups defined by a factor such as disease status, disease subtype, or geographical region, and identifying the differential abundant (DA) taxa associated with the groups. The taxa-group relationships may be confounded by covariates producing spurious associations [Vujkovic-Cvijin et al., 2020]. Of primary interest are latent differential status variables $\bar{h}_j$, defined as $\bar{h}_j = 2$ ($\bar{h}_j = 1$) if taxon $j$ is (not) DA. The set of DA taxa is $\emptyset = \{j : \bar{h}_j = 2, j = 1, \ldots, p\}$.

For $n$ samples and $p$ taxa, the data are arranged in an $n \times p$ abundance matrix with each matrix element representing the observed count or abundance of a taxon (matrix column) in a study sample or subject (matrix row). Usually, there are far fewer samples than taxa, i.e., $n < < p$. We write the taxa abundance matrix as $Z = ((Z_{ij}))$, where $Z_{ij}$ is the number of reads of taxon $j$ in sample $i$, and $Z_i = (Z_{i1}, \ldots, Z_{ip})'$ is the vector of taxa abundances for sample $i$. Suppose there are $T$ covariates, $X_i$, and the $n$ by $T$ covariate matrix is denoted by $X = ((X_{it}))$. With $L_i$ representing the observed sequencing depth or library size, the relative abundance vector of the $i$th subject, denoted by $q_i = (\hat{q}_{i1}, \ldots, \hat{q}_{ip})'$, equals $Z_i/L_i$. For $K \geq 2$ groups, the group membership of the $i$th sample is denoted by $k_i$, and there are $n_k$ samples belonging to the $k$th group, $k = 1, \ldots, K$. Now suppose there are two samples, $i_1$ and $i_2$, belonging to different groups ($k_{i_1} \neq k_{i_2}$) but having similar covariates ($X_{i_1} \approx X_{i_2}$). Then, by the intuitive characterization of differential statuses in Section 1, a taxon $j$ is non-DA if $\hat{q}_{i_1j} \approx \hat{q}_{i_2j}$. However, if $\hat{q}_{i_1j}$ very different from $\hat{q}_{i_2j}$, then taxon $j$ is DA.

As discussed in Section 1, biological zeros refer to the complete absence of a taxon in an entire (say, $k$th) group and manifest as a matrix column of zeros for all $n_k$ subjects. These
taxa are easily identified as DA and removed from the dataset in a pre-processing step. On the other hand, technical zeros are caused by processing or sequencing batch effects and sampling zeros are randomly missing taxa in specific samples. Distinguishing and appropriately accounting for these “non-biological” zeros in differential analysis requires more sophisticated techniques.

The proposed ZIBNP model postulates a two-component mixture for each element of $Z$, namely, a point mass at zero representing technical zeros and a Bayesian nonparametric (BNP) model, denoted by $\mathcal{F}$, under which sampling zeros may stochastically occur. More formally, with $\mathcal{F}_{ij}$ denoting the marginal distribution of $Z_{ij}$ under BNP model $\mathcal{F}$ and $I_{\{0\}}$ representing a point mass at 0:

$$Z_{ij} \mid r_{ij}, \mathcal{F}_{ij} \overset{\text{iid}}{\sim} r_{ij}I_{\{0\}} + (1 - r_{ij})\mathcal{F}_{ij}, \quad i = 1, \ldots, n; \; j = 1, \ldots, p,$$

where $r_{ij}$ is the probability of a technical zero in sample $i$ and taxon $j$. Let $R = ((r_{ij}))$ denote the $n \times p$ matrix of technical zero probabilities. The first mixture component represents technical zeros and the second component includes sampling zeros. Section 2.1 develops BNP model $\mathcal{F}$. Section 2.2 describes the model for technical zeros.

### 2.1 Bayesian nonparametric (BNP) model

The BNP model $\mathcal{F}$ assumes the following distribution for the taxa abundances:

$$Z_i \mid q_i \overset{\text{iid}}{\sim} \text{Multinomial}(L_i, q_i), \quad i = 1, \ldots, n,$$

where probability $p$-tuple $q_i = (q_{i1}, \ldots, q_{ip})'$ is a less noisy version of relative abundance vector $\hat{q}_i = (\hat{q}_{i1}, \ldots, \hat{q}_{ip})'$ and forms the $i$th row of an $n$ by $p$ matrix, $Q = ((q_{ij}))$. Let $\tilde{q}_j$ denote the $j$th matrix column, so that $Q = [\tilde{q}_1, \ldots, \tilde{q}_p]$. The conditional probability of a sampling zero is

$$P[Z_{ij} = 0 \mid Q, \mathcal{F}] = (1 - q_{ij})^{L_i}. \quad (3)$$

### Dimension reduction by unsupervised clustering

One of the challenges of microbiome data analysis is the “small $n$, large $p$” setting causes severe collinearity in the relative
abundance matrix and gives inefficient inferences. BNP model $\mathcal{F}$ resolves these issues by utilizing lower-dimensional matrix structure. Specifically, unsupervised clustering of the $p$ taxa allocates them to $H$ latent clusters, where $H$ is a priori unknown. Each cluster is characterized by a $n$-variate motif delineating the relative abundance pattern shared by all taxa belonging to the cluster. Dimension reduction is achieved when $H$ is less than $p$.

From a mathematical standpoint, latent clusters are unavoidable in small $n$, large $p$ matrices because the matrix rank cannot exceed $n$, and this produces redundancies in the large number of matrix columns. From a scientific standpoint as well, biomarkers belonging to common phylogenetic groups or functional pathways tend to be highly correlated [e.g., Lee et al., 2020]. The phenomenon of biomarker clusters having similar patterns in high-dimensional genomic, epigenomic, and transcriptomic datasets is well-documented and has been utilized to achieve dimension reduction [e.g., Medvedovic et al., 2004, Kim et al., 2006, Dunson et al., 2008, Dunson and Park, 2008, Guha et al., 2022, Gu et al., 2023].

More formally, define taxon-to-cluster mapping variables, $c_1, \ldots, c_p$, with the random event $\{c_j = u\}$ indicating that the $j$th taxon is allocated to the $u$th latent cluster, $u = 1, \ldots, H$. Let $m_u = \sum_{j=1}^{p} I(c_j = u)$ be the number of taxa in the $u$th latent cluster. Mapping vector $c = (c_1, \ldots, c_p)$ is given a Chinese restaurant process (CRP) prior with a positive precision or mass parameter $\alpha$ [Müller and Mitra, 2013]. With $[\cdot]$ representing densities with respect to a dominating measure, we have

$$[c | \alpha] = \frac{\Gamma(\alpha)^H \prod_{u=1}^{H} \Gamma(m_u)}{\Gamma(\alpha + p)}, \quad c \in \mathcal{P}_p,$$

where $\mathcal{P}_p$ is the set of all partitions of $p$ taxa into one or more latent clusters. CRPs achieve dimension reduction in the large number of taxa because the random number of clusters, $H$, is asymptotically equivalent to $\alpha \log(p)$ as $p \to \infty$ [Lijoi and Prünster, 2010], so that typically $H \ll p$.

**Cluster motifs** Motif $q_u^* = (q_{1u}^*, \ldots, q_{nu}^*)'$ embodies the “signal” underlying the relative abundances of the $u$th cluster’s member taxa. In other words, the $H$ motifs represent the across-sample pattern shared by all taxa belonging to a latent cluster. Expression (2)
establishes that relative abundance vectors $\tilde{q}_1, \ldots, \tilde{q}_p$ are just noisy versions of the corresponding matrix $Q$ columns, $q_1, \ldots, q_p$. This suggests that the matrix columns $Q$ are identical to their motifs:

$$\tilde{q}_j = q^*_c, \quad \text{for taxa } j = 1, \ldots, p. \quad (5)$$

Because the measured taxa abundances in (2) are random, the taxa belonging to a cluster have similar, although not necessarily identical, relative abundances.

Conditional on the mapping vector $c = (c_1, \ldots, c_p)$, matrix $Q$ is fully determined by the lower-dimensional matrix $Q^* = ((q^*_{iu}))$ of dimension $n$ by $H$, where $H \ll p$. Additionally, since $Q$ is row-stochastic, each row of matrix $Q^*$ satisfies

$$\sum_{u=1}^{H} m_u q^*_{iu} = 1, \quad i = 1, \ldots, n, \quad (6)$$

implying that the cluster motifs are linearly dependent. Furthermore, applying equation (5) and the informal characterization of differential taxa in Section 1, we find that all taxa in a latent cluster must have identical differential statuses. That is, a cluster is collectively DA or non-DA. Consequently, analogously to the $p$ taxa differential status variables, we can define cluster differential status variables $h_1, \ldots, h_H$ and describe the set of DA taxa equivalently as

$$\mathcal{D} = \{j : h_c = 2, \quad j = 1, \ldots, p\}. \quad (7)$$

**Incorporating covariates**

To model the regression relationships of the cluster motifs, we begin by selecting a non-DA reference cluster. Strategies for choosing this special non-DA cluster are discussed in the sequel. Because the CRP is symmetric with respect to the cluster labels, the reference cluster is labeled $u = 1$ without loss of generality. We assume

$$\eta_{iu} = \log\left(\frac{q^*_{iu}}{q^*_{i1}}\right) \overset{\text{indep}}{\sim} N\left(\beta_{0k,u} + \sum_{l=1}^{T} x_{il} \beta_{lk,u}, \sigma_k^2\right), \quad (8)$$

for $i = 1, \ldots, n$, and $u > 1$. 

9
The $q_{iu}^*$ are identifiable due to constraint (6) and matrix $Q$ is then available from (5). We assume that $\sigma_e^2$ is small enough that most of the variability in the $\eta_{iu}$’s is captured by the regression means. As a result, in the (unrealistic) situation in which there are no covariates (i.e., $x_{il} \equiv 0$), the latent elements $q_{iu}^*$ of all $n_k$ samples in any group $k$ are approximately equal.

**Differential statuses of clusters and taxa** By design, the reference cluster (taxon) is non-DA, and the differential status variable for the cluster is $h_1 = 1$. For clusters $u > 1$, since the regression vectors $\beta_1, \ldots, \beta_K$ adjust for the covariates, the cluster differential statuses are available as

$$h_u = \begin{cases} 
1 & \text{if } \beta_1 = \cdots = \beta_K, \\
2 & \text{otherwise.} 
\end{cases}$$

(9)

In other words, a non-reference cluster is non-DA if and only if its $K$ regression vectors are identical. Further, the differential statuses of the taxa are identical as their parent clusters. Returning to a hypothetical no-covariate dataset, for any non-DA cluster $u$, expressions (8) and (9) show that the latent elements $q_{iu}^*$ of all $n$ samples, irrespective of their groups, are approximately equal. In the presence of covariates, the differential statuses may not be obvious from the sample patterns of $q_{iu}^*$ or the noisier relative abundances, but can be inferred using (9).

Evidently, a reasonable prior must allow multiple regression vectors to be equal with positive probability. An appropriate prior can be specified as follows. Let $M$ be a positive integer, $1_M$ be the column vector of $M$ ones, and $X^\dagger = [1_n : X]$ be the extended covariate matrix of dimension $n$ by $(T + 1)$. Let $\delta_\mu$ denote a point mass at $\mu$. For a parameter vector $\gamma$ with positive elements, $\text{Dir}_M(\gamma)$ represents a Dirichlet distribution in $\mathcal{R}^M$. The
regression vectors have the $M$-component mixture prior:

$$
\beta_{ku} \overset{\text{i.i.d.}}{\sim} \sum_{m=1}^{M} \pi_m \delta_{\mu_m}, \quad k = 1, \ldots, K, \text{ and } u > 1, \quad \text{(10)}
$$

$$
\pi = (\pi_1, \ldots, \pi_M)^\prime \sim \text{Dir}_M(\frac{\alpha_0}{M} \mathbf{1}_M), \quad \alpha_0 > 0,
$$

$$
\mu_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{T+1}(0, \tau^2 (X^T X)^{-1}), \quad m = 1, \ldots, M, \quad \text{(11)}
$$

and where $\tau^2$ follows an inverse-gamma hyperprior with parameters $a_\tau$ and $b_\tau$. The finite mixture specification is key because it allows ties among the $\beta_{ku}$. In simulation studies and data analyses, we have found that $M$ equal to 5 or 6 gives satisfactory results.

These ideas are demonstrated by a toy example in Figure 1, where we have $K = 2$ groups of subjects (i.e., adults and children), $p = 50$ taxa, and several subject-specific covariates. The taxa are allocated to $H = 7$ clusters. In the figure, row vector $X_i^\dagger$ denotes the $i$th row of matrix $X^\dagger$ of length $(T+1)$ and $\epsilon_{iu} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$. The $K(H-1) = 12$ vectors, each consisting of the $(T+1)$ regression coefficients for a group–non-reference cluster combination, arise from a finite mixture model with $M = 5$ multivariate components. For $u > 1$ (i.e., non-reference clusters), the colors represent the five-mixture model components from which each $\beta_{ku}$ vector is drawn in (10). For example, the regression vectors $\beta_{14}$, $\beta_{24}$, and $\beta_{27}$ are identical because they are drawn from the same mixture component (shown in blue). Applying criterion (9), cluster 4 is non-DA, and clusters 2, 3, and 7 are DA in adults and children. All taxa have the same differential status as their parent cluster. Specifically, taxa 4, 5, 7, and 9, which belong to cluster 4, are all non-DA; taxa 2 and 8, which belong to cluster 3, are both DA.

**Choosing the reference cluster** For correctly calling DA clusters using criterion (9), it is important to choose a non-DA cluster as the reference. This can be seen by the following example. For simplicity, imagine a no-covariate dataset with $K = 2$ groups, $H = 3$ latent clusters, and small $\sigma_e^2$ in (8). Suppose we pick a reference cluster ($u = 1$) for which $q_{i1}^* = 0.5$ and $q_{i2}^* = 0.1$ for all samples $i_1$ and $i_2$ for which $k_{i1} = 1$ and $k_{i2} = 2$. That is, a DA cluster was mistakenly chosen to be the reference. Now, let cluster $u = 2$ be actually non-DA with $q_{i2}^* = 0.1$ for all $i$. Since $\sigma_e^2$ is small, relation (8) gives $\beta_{012} \approx -\log 5$.
Figure 1: Cartoon illustration of the model-based procedure for calling the differential statuses of the taxa. There are $K = 2$ subject groups (adults and children), $p = 50$ taxa, multiple subject-specific covariates, and $H = 7$ latent clusters. The regression vectors for the non-reference clusters ($u > 1$) are shown. The different colors identify the corresponding mixture components from which each $\beta_{ku}$ is drawn. Applying criterion (9), all taxa belonging to clusters 2, 3 and 7 are DA, and all taxa belonging to cluster 4 are non-DA. See the text for details.
and $\beta_{022} \approx 0$, and so, (9) wrongly classifies cluster 2 as DA. Let cluster $u = 3$ be actually DA with $q_{i1}^* = 0.25$ and $q_{i2}^* = 0.05$ whenever $k_{i1} = 1$ and $k_{i2} = 2$. Then $\beta_{0k3} \approx -\log 2$ for $k = 1, 2$, and again, (9) misclassifies cluster 3 as non-DA. These errors are avoided if non-DA cluster 2 is designated as the reference.

The following are some practical strategies for choosing an appropriate reference cluster. We have used the second option in the simulation study and data analyses.

- **Minimum variance cluster**  Adapting the technique of Nearing et al. [2022], we fix the taxon for which the relative abundances are least variable as the (singleton) reference cluster.

- **Artificial reference taxon**  Augment abundance matrix $Z$ with an artificial taxon with unit abundance counts for all $n$ samples. Since the taxa are exchangeable, label the taxon as $j = 1$ and the remaining (actual) taxa as $j = 2, \ldots, (p + 1)$. Increment all the sampling depths, $L_1, \ldots, L_n$, by 1 to accommodate the additional count per matrix row. Assume that the artificial taxon constitutes its own cluster, and so there are $(H + 1)$ clusters in the augmented dataset. Since the sampling depths of microbiome datasets are many orders of magnitude greater than 1, this additional “taxon” has small relative abundances, is therefore guaranteed to be non-DA, and has virtually no effect on the eventual inferences.

Irrespective of the manner in which the reference cluster is chosen, the clusters are reordered so that the reference has the label $u = 1$.

### 2.2 Modeling non-biological zeros

Although the aforementioned aspects of the ZIBNP model are able to accommodate covariates, high-dimensionality, and compositionality, the challenge of high sparsity still remains. As discussed, biological zeros are detected in a straightforward manner. Non-biological (i.e., sampling or technical) zeros require a more nuanced approach. Specifically, BNP submodel $F$ accounts for sampling zeros, which are therefore informative about the differential statuses. By contrast, technical zeros are caused by random sequencing errors that
obfuscate important aspects of $F$. We cast technical zeros in a missing data framework that identifies the source of non-biological sparsity to more accurately call the differential taxa.

Returning to the first mixture component in expression (1), we begin by modeling $r_{ij}$, the probability of technical zeros in sample $i$ and taxon $j$. The log-sampling depths and covariates are known predictors of technical zeros [Ma et al., 2020, Jiang et al., 2021]. For a cluster-specific random effects vector, $\lambda_{iu} = (\lambda_{0iu}, \ldots, \lambda_{T+1,iu})'$, we therefore assume

$$\text{logit}(r_{ij}) = \lambda_{0icj} + \sum_{t=1}^{T} \lambda_{ticj} x_{it} + \lambda_{T+1,icj} \log(L_i), \; i = 1, \ldots, n,$$  \hspace{1cm} (12)$$

where $\lambda_{tiiu} \sim \text{i.i.d. } N(0, \tau_\lambda^2)$ for all $t$, $i$, and $u$. Denote by $\Lambda$ the collection of cluster-specific random matrices $\Lambda_1, \ldots, \Lambda_H$, where $\Lambda_u = ((\lambda_{tiiu}))$ is the $n \times (T + 2)$ matrix of regression coefficients for the $u$th cluster in equation (12). Because all $m_u$ taxa belonging to the $u$th cluster have the same probability of technical zeros for a given sample $i$, we write

$$r_{icj}^* = r_{ij}, \; j = 1, \ldots, p.$$  

From expressions (1) and (3), and applying the CRP taxon-to-cluster mappings, the probability of a non-biological (technical or sampling) zero for taxon $j$ of sample $i$ is then

$$P(Z_{ij} = 0 \mid c_j, \Lambda, Q^*) = r_{icj}^* + (1 - r_{icj}^*)(1 - q_{icj}^*)L_i.$$  \hspace{1cm} (13)$$

The lower-dimensional framework of submodel $F$ and equation (12) imply that the taxa in a cluster display somewhat similar zero patterns. The next innovation helps distinguish sampling zeros from technical zeros.

**Censoring framework for technical zeros** We interpret expression (1) as an independent censoring mechanism that converts the true abundance count, $\tilde{Z}_{ij}$, arising from submodel $F$, to an observed technical zero with probability $r_{icj}^*$. We denote the true, partially latent abundance matrix arising from $F$ by $\tilde{Z} = ((\tilde{Z}_{ij}))$. Parameters directly related to the DA taxa then rely on matrix $\tilde{Z}$ rather than data matrix $Z$. Let $\delta_{ij}$ represent the indicator that $\tilde{Z}_{ij}$ is uncensored, so that $\delta_{ij} \sim \text{Bernoulli}(1 - r_{icj}^*)$. The observed abundance
counts are related to the true counts as

\[ Z_{ij} = \begin{cases} \tilde{Z}_{ij} & \text{if } \delta_{ij} = 1, \\ 0 & \text{if } \delta_{ij} = 0. \end{cases} \]

If \( Z_{ij} > 0 \), then clearly, \( \delta_{ij} = 1 \). On the other hand, \( Z_{ij} = 0 \) corresponds to either a sampling zero (\( \delta_{ij} = 1 \)) or technical zero (\( \delta_{ij} = 0 \)).

Let latent set \( J_i = \{ j : \delta_{ij} = 0, j = 1, \ldots, p \} \) comprise taxa with censored counts in the \( i \)th sample. Using latent matrix \( \tilde{Z} \), the observed sampling depths are \( L_i = \sum_{j \not\in J_i} \tilde{Z}_{ij} \). The unobserved sampling depth relies only on the censored taxa: \( \tilde{S}_i = \sum_{j \in J_i} \tilde{Z}_{ij} \). The true sampling depth is \( \tilde{L}_i = \sum_{j=1}^p \tilde{Z}_{ij} \) and equals \( (L_i + \tilde{S}_i) \).

All aforementioned aspects of the ZIBNP model, from equations (1) to (13), are amended by replacing the observed matrix \( Z \), count \( Z_{ij} \), and sampling depth \( L_i \), wherever they occur, with their true (possibly latent) counterparts, i.e., true abundance matrix \( \tilde{Z} \), true abundance counts \( \tilde{Z}_{ij} \), and true sampling depth \( \tilde{L}_i \). Unlike traditional approaches for censored outcomes, however, censoring indicator \( \delta_{ij} \) is unknown for the zero abundances, and the probability of a technical zero is

\[ P[\delta_{ij} = 0 \mid Z_{ij} = 0, c_j, \tilde{L}_i, \Lambda, Q^*] = \frac{r^*_{ic_j}}{r^*_{ic_j} + (1 - r^*_{ic_j})(1 - q^*_{ic_j})L_i}, \quad (14) \]

with the conditional probability of a sampling zero given by the complementary event. Since the sampling depths are usually large, a technical zero is much more likely than a sampling zero, unless \( q^*_{ic_j} \) is very small. In any case, equation (14) is applied to call the technical and sampling zeros (i.e., a posteriori generate the unknown censoring indicators) by the MCMC procedure outlined in Section 3.1.

As mentioned, all inferences about parameters related to differential analysis rely on the partially latent, true abundance matrix \( \tilde{Z} \). The following result, whose proof appears in Supplementary Material, facilitates posterior inferences about the true abundance matrix \( \tilde{Z} \).

**Theorem 2.1.** Suppose the censoring indicators \( \delta_{ij} \) corresponding to the zero abundances are known. For sample \( i = 1, \ldots, n \), let \( \tilde{q}_i = \sum_{j \in J_i} q^*_{ic_j} \). Then
1. True sampling depth \( \tilde{L}_i \) has a negative binomial distribution:

\[
\tilde{L}_i \mid J_i, c, Q^* \sim \text{NegBin}(L_i, 1 - \tilde{q}_i),
\]

for which \( \tilde{L}_i \) is the random number of i.i.d. Bernoulli trials with success probability \((1 - \tilde{q}_i)\) and \(L_i\) is the prespecified number of successes. Hence, \( \tilde{S}_i = \tilde{L}_i - L_i \).

2. Let the vector of \(|J_i|\) unobserved taxa abundances be \( \tilde{Z}_i^{(0)} = (\tilde{Z}_{ij} : \delta_{ij} = 0, j = 1, \ldots, p) \). Define \( w_{ij} = q_{icj}/\tilde{q}_i \) for \( j \in J_i \), and probability vector \( w_i = (w_{ij} : j \in J_i) \) of length \(|J_i|\). Then

\[
\tilde{Z}_i^{(0)} \mid \tilde{L}_i, J_i, c, Q^* \sim \text{Multinomial}(\tilde{S}_i, w_i),
\]

where \( \tilde{S}_i = \tilde{L}_i - L_i \).

Finally, we complete the ZIBNP model by assigning standard conjugate priors to the remaining hyperparameters. Figure 2 presents a directed acyclic graph (DAG) of the parameters.

3 Posterior inferences

Section 3.1 outlines the Monte Carlo inference procedure. Section 3.2 describes the technique for detecting the DA taxa.

3.1 MCMC algorithm

After initialization by naive estimation techniques, as described in Supplementary Material, the model parameters are iteratively updated using MCMC procedures. All parameters except the \( \eta_{iu} \)'s are updated by Gibbs sampling. Although the full conditional of \( \eta_{iu} \) does not have a closed form, it is log-concave, and \( \eta_{iu} \) can therefore be generated by adaptive rejection sampling [Gilks and Wild, 1992]. See Supplementary Material for an outline of the MCMC steps. The post-burn-in MCMC sample of the parameters (including mapping variables, group-cluster regression coefficients, censoring indicators, true abundance counts with imputed technical zeros, and differential statuses) is stored and processed for posterior inferences.
Figure 2: DAG representation of the ZIBNP model highlighting parameters related to handling sparsity (“zero inflation”), high-dimensionality (“clustering”), and DA taxa detection (“differential abundance”). Circles represent stochastic parameters, solid rectangles represent the data and deterministic variables, and open rectangles represent prespecified constants.
3.2 Calling taxa differential statuses

Using $L$ post-burn-in MCMC samples, a straightforward estimator of $P[\tilde{h}_j = 1 \mid \mathbf{Z}]$, the posterior probability that taxon $j$ is non-DA, is $\sum_{l=1}^{L} I(\tilde{h}_j^{(l)} = 1)/L$. However, the following theoretical result gives a more precise “Rao-Blackwellized” estimator that is especially accurate for taxa with ambiguous differential statuses. But first, for group $k$ and non-reference latent cluster $u > 1$, let membership variable $v_{ku}$ document the $M$-mixture component (10) from which regression vector $\beta_{ku}$ is drawn. That is,

$$\beta_{ku} = \mu_{v_{ku}}, \quad k = 1, \ldots, K, \; u = 2, \ldots, H.$$  

(15)

The theorem follows from the law of total probability. We omit the proof for brevity.

**Theorem 3.1.** Let the set $\Theta^-$ contain all ZIBNP model parameters except the mapping variables, differential status variables for the taxa and clusters, membership variables (15), and any parameters having a deterministic relationship with these variables. Let $P_*[\cdot]$ denote the conditional probability, $P[\cdot \mid \Theta^-, H, \mathbf{Z}]$. Then

1. For non-reference cluster $u$, the conditional posterior probability that a cluster is non-DA is

$$P_*[h_u = 1] = \sum_{m=1}^{M} \prod_{k=1}^{K} P_*[v_{ku} = m], \quad u = 2, \ldots, H.$$ 

2. For $j = 1, \ldots, p$, the conditional posterior probability that the $j$th taxon is non-DA is

$$P_*[\tilde{h}_j = 1] = P_*[c_j = 1] + \sum_{c=2}^{H} P_*[h_u = 1]P_*[c_j = u].$$

Applying Theorem 3.1, an estimator for the posterior probability that the $j$th taxon is non-DA is then

$$\hat{P}[\tilde{h}_j = 1 \mid \mathbf{Z}] = \frac{1}{L} \sum_{l=1}^{L} P[\tilde{h}_j = 1 \mid \Theta^-_l, H(l), \mathbf{Z}],$$

where $\Theta^-_l$ and $H(l)$ are the parameter values generated at the $l$th MCMC iteration. Apply the Rao-Blackwell theorem, it is easy to verify that this estimator is more precise than $\sum_{l=1}^{L} I(\tilde{h}_j^{(l)} = 1)/L$ even though both estimators are consistent as $L \to \infty$. 

18
The procedure is applied to infer the posterior probabilities of the taxa differential statuses along with uncertainty estimates. Given a nominal FDR (say, 5%), we apply the Bayesian FDR approach of Newton et al. [2004] to choose an appropriate threshold, \( \kappa \), for calling the DA taxa. An estimate of differential status variable \( \tilde{h}_j \) is

\[
\hat{h}_j = 1 + \mathcal{I}\left( \hat{P}[\hat{h}_j = 1 \mid \mathbf{Z}] \geq \kappa \right), \quad j = 1, \ldots, p,
\]

where \( \mathcal{I}(\cdot) \) represents the indicator function. This gives the estimated set of DA taxa, \( \hat{\mathcal{D}} = \{ j : \hat{h}_j = 2, \ j = 1, \ldots, p \} \).

4 Simulation studies

We investigated the accuracy of the proposed ZIBNP approach using artificial microbiome datasets with different sparsity levels. For comparison, we analyzed the data using existing statistical methods with available R packages capable of detecting DA taxa in the presence of covariates.

Data generation  The following steps, which differ in several important aspects from the ZIBNP model, were used to generate the artificial datasets. For \( n = 100 \) subjects belonging to \( K = 2 \) groups, we generated abundance matrix \( \mathbf{Z} \) for \( p = 1,000 \) taxa, with taxon 1 constituting an artificial reference cluster with unit abundance counts. For four sparsity levels modulated by a simulation parameter \( \lambda_0 \) in Step (f) below, we independently generated 25 datasets as follows:

(a) Covariates  We used the actual covariates of a publicly available oral microbiome dataset [Burcham et al., 2020] consisting of adults and children as the subject groups. We selected \( T = 4 \) covariates, namely, three binary covariates (indicators of antibiotics intake during the last 6 months, adequate brushing habits, and sex) and one continuous covariate (BMI). Randomly sampling the covariates of 50 subjects, we replicated the same covariates in adults and children to obtain the covariate matrix,
$X$, of dimension 100 by 4, of the artificial dataset. That is, if $X^{(0)}$ denotes the $50 \times 4$ covariate matrix randomly sampled from the oral microbiome dataset, then

$$X = \begin{bmatrix} X^{(0)} \\ X^{(0)} \end{bmatrix},$$

where the first 50 subjects represent adults and the remaining 50 subjects represent children in the artificial dataset. Although unnecessary for the downstream analysis, this structure helps informally validate the detected DA taxa, as discussed later.

(b) **Mapping variables** We generated the true number of clusters, $H$, from a discrete uniform distribution on \{8, 9, ..., 20\}. Applying the `rpartitions` package in R, we generated a $H$-partition of $p$ objects using the `rand_partitions` function of the package. The true taxon-to-cluster mapping variables, $c_1, \ldots, c_p$, were set equal to this randomly generated partition and therefore relied on a very different stochastic mechanism than CRP (4).

(c) **True differential statuses** In relation (10), setting $M = 7$ and $\tau^2 = 1$, and since $(T + 1) = 5$, the finite mixture locations, $\mu_m$, were independently generated from distribution $N_5(0, (X^T X)^{-1})$, for $m = 1, \ldots, 7$. The mixture probability vector $\pi$ was generated from the Dirichlet distribution, $\text{Dir}_7(\frac{1}{7} 1_7)$. For the $(H - 1)$ non-reference clusters, the $K(H - 1) = 2(H - 1)$ regression vectors, $\beta_{ku}$, were generated from finite mixture model (10) consisting of $M = 7$ components. The true differential statuses $h_2, \ldots, h_H$ of the non-reference clusters were calculated by applying (9). As usual, $h_1 = 1$ for the reference cluster. The true set of DA taxa, $D$, was evaluated using (7).

(d) **Probability matrix $Q$** Unlike the ZIBNP model, in order to generate the motif elements of the artificial datasets, we assumed

$$q_{iu}^* = \alpha_{ih_u} \exp(\zeta_{iu}), \quad u > 1,$$

where $\alpha_{ih_u}$ is the subject-specific proportionality constant that depends on the DA
status of cluster $u$, and the $\zeta_{iu}$’s were generated as

$$\zeta_{iu} \overset{\text{indep}}{\sim} N\left(\beta_{0k_iu} + \sum_{l=1}^{T} x_{il} \beta_{lk_iu}, \sigma_{zi}^2\right), \quad i = 1, \ldots, n, \text{ and } u > 1,$$

with $\sigma_{zi}^2$ chosen so that $R^2$ exceeded 99%. Since microbiome data are compositional,

$$\sum_{i=1}^{p} q_{ij} = \sum_{u=1}^{H} m_{iu} q_{iu}^{*} = 1,$$  

so that

$$\sum_{u:h_{u}=1} \alpha_{i1} m_{iu} \exp(\zeta_{iu}) + \sum_{u:h_{u}=2} \alpha_{i2} m_{iu} \exp(\zeta_{iu}) = 1.$$  

For $h = 1, 2$, let $\rho_{ih} = \sum_{u:h_{u}=h} \alpha_{ih} m_{iu} \exp(\zeta_{iu})$ be the total probability assigned to the non-DA and DA taxa, respectively, in equation (19). We generated $\rho_{i1} \sim \text{beta}(0.5A, 0.5A)$, for $A$ large, and set $\rho_{i2} = 1 - \rho_{i1}$. Thereby, we computed

$$\alpha_{ih} = \frac{\rho_{ih}}{\sum_{u:h_{u}=h} \alpha_{ih} m_{iu} \exp(\zeta_{iu})}, \quad h = 1, 2,$$

(20)

(21)

to evaluate the matrix $Q^*$. Then, applying equation (5), we obtained probability matrix $Q$.

(e) **Uncensored taxa abundances**  To mimic the range of sampling depths observed in actual microbiome datasets, we generated the true sampling depths $\tilde{L}_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(10,000) \times \text{Poisson}(100)$, for subjects $i = 1, \ldots, 100$. The true taxa abundances were generated as $\tilde{Z}_i \sim \text{Multinomial}(\tilde{L}_i, q_i)$.  

(f) **Observed taxa abundances**  Setting $\lambda_{iu} = \lambda_0 1$, where $\lambda_0 \in \mathcal{R}$, and using the true sampling depths $\tilde{L}_i$, we applied equation (12) to compute the probability of technical zeros, $r^*_{iu}$, for $u = 2, \ldots, H$. We generate the censoring indicators as $\delta_{ij} \overset{\text{indep}}{\sim} \text{Bernoulli}(1-r^*_{ij})$, for $j = 1, \ldots, p$. Thereafter, we calculated the taxa abundance matrix $Z$ using (14). Varying simulation parameter $\lambda_0$ in the set $\{-0.1, -0.059, -0.001, 0.023\}$ produced a range of sparsity levels matching typical microbiome data; see Table 1.

Figure 3 displays a heat map of the true log ratios, $\zeta_{iu}$, for the $H = 22$ true clusters of a randomly chosen dataset. The $p = 1,000$ taxa are not shown in the figure. Due to special
Figure 3: For a randomly chosen artificial dataset with $H = 22$ clusters, heat map of the true log ratios, $\zeta_{iu}$, for subject $i$ (rows) and cluster $u$ (columns), underlying the generated taxa abundances. The first 50 subjects are adults (group 1) and the remaining 50 subjects are children (group 2). The horizontal line separates the two groups of subjects. The column labels in the lower axis have the format “$u_{hu}$” for $u = 1, \ldots, H$, to indicate the index and true differential status of the $u$th cluster. Reference cluster 1 has column label “1_1” and is non-DA by design. See Section 4 for further discussion.
structure (16), motif elements (17), and relatively small $\sigma^2_\xi$ in (18), we can accurately call the true differential statuses of the non-reference clusters by a visual comparison of the color patterns of the two groups in Figure 3. Specifically, for $u = 1, \ldots, H$, the $u$th cluster, represented by the $u$th column in Figure 3, is non-DA if the color patterns in the upper and lower column halves are identical, in which case the column label is “$u.1$”; otherwise, the cluster is DA and its column label is “$u.2$”. For example, clusters 1 and 2 are non-DA ($h_1 = h_2 = 1$) because the patterns are identical for adults and children. Cluster 3 is DA ($h_3 = 2$).

Disregarding knowledge of the generation mechanism, the Section 3 procedure was applied to analyze each artificial dataset and make posterior inferences using the ZIBNP model. Applying the Section 3.2 strategy, we post-processed the MCMC sample to estimate the DA posterior probabilities, $P[\hat{h}_j = 2 \mid Z]$. Varying the posterior probability thresholds for calling the DA taxa, we compared $\hat{\mathcal{D}}$, the detected DA taxa using that threshold, with the true $\mathcal{D}$; see Step (c). The sensitivities and specificities over the range of thresholds were evaluated to produce the ROC curve for each dataset. Additionally, as described in Section 3.2, assuming a nominal FDR of 5%, we applied the Bayesian FDR procedure to call the DA taxa, and thereby, evaluated the FDR and sensitivity achieved by ZIBNP for each simulation dataset.

Using area under the curve (AUC), false discovery rate (FDR), and sensitivity as the evaluation criteria, we compared our technique with some well-established methods for

| $\lambda_0$ | % zeros |
|------------|---------|
| -0.100     | 13%     |
| -0.059     | 25%     |
| -0.001     | 50%     |
| 0.023      | 60%     |

Table 1: Averaging over the 25 artificial datasets, observed sparsity in abundance matrix $Z$ for different $\lambda_0$ values.
differential abundance analysis with covariates, focusing only on methods implemented in publicly available R packages:

(i) ANCOMBC [version 1.0.5; Lin and Peddada, 2020]: We used the \texttt{ancombc} function with default settings. The function fits compositional data by making a log transformation and using a sample-specific offset term for bias correction.

(ii) Maaslin2 [version 1.4.0; Mallick et al., 2021]: The \texttt{Maaslin2} function was used with total sum scaling normalization and covariate fixed effects. This technique fits a generalized linear model to each feature abundance with respect to the covariates, checking for significance using the Wald test.

(iii) Metagenomeseq [version 1.32.0; Paulson et al., 2013]: Function \texttt{fitZIG} was applied to fit a zero-inflated Gaussian version of the technique. The workflow involved normalizing the data using cumulative sum scaling with default settings.

(iv) DESeq2 [version 1.30.1; Love et al., 2014]: We used the \texttt{DESeq2} function with argument \texttt{sizefactor} set to “poscounts.” The R function performs a likelihood ratio test for a reduced model containing all covariates except the grouping factor.

The FDRs of the competing statistical methods were controlled via taxa-specific q-values, i.e., adjusted p-values [Benjamini and Hochberg, 1995]. Using the q-values, a method’s sensitivities and specificities for critical values between 0 to 1 produced the ROC curve for each dataset. Additionally, for a target FDR of 5%, the DA taxa were called using the Benjamini-Hochberg procedure to evaluate a method’s achieved FDR and sensitivity for each dataset.

Averaging over the 25 datasets, Figure 4 displays the ROC curves of all the methods, with the panels corresponding to the four sparsity levels in Step (f). The corresponding average AUCs, along with 95% confidence intervals for AUC, are presented in Table 2. Even though the accuracy of ZIBNP deteriorated with increasing sparsity as expected, we find that ZIBNP has a substantially higher AUCs than the competing methods for all sparsity levels.
Figure 4: For the simulation study, ROC plots comparing the methods ZIBNP, ANCOMBC, fitZIG, and MaAsLin2. The panels represent different sparsity levels ranging from 13% to 60% zeros.

For 13% and 25% sparsity, Figures 5 and 6 respectively represent the AUC, FDR, and sensitivity for the statistical methods. The method DESeq2 declared all taxa as non-DA in the simulated datasets, and is not shown in the plots. In both plots, ANCOMBC had a median AUC of around 0.5 (sub-figure A), exhibited relatively high FDR (sub-figure B) and low sensitivity (sub-figure C). For fitZIG, the typical AUC in sub-figure A was below 0.5 with an elevated FDR in sub-figure B, but its sensitivity was relatively high (sub-figure...
Figure 5: For 13% zeros in the abundance matrix, summarizing over the 25 artificial datasets of Section 4, boxplots of AUC (A), FDR (B), and sensitivity (C) for the different statistical methods. The dashed horizontal line in sub-figure B indicates the target FDR of 5% for all the methods.

C). The AUCs for MaAsLin2 were relatively high and the FDRs were well below the target FDR, but the sensitivity was somewhat low. Further investigation into the performance of MaAsLin2 revealed that for most of the 25 artificial datasets and all sparsity levels, the technique failed to detect most of the DA taxa. These results are consistent with previous studies that found similar AUCs and FDRs for fitZIG and MaAsLin2 [e.g., Thorsen et al., 2016, Hawinkel et al., 2017, Mallick et al., 2021].

By contrast, the figures and tables reveal that ZIBNP displayed substantially higher AUC, well-controlled FDR, and high sensitivity. In summary, ZIBNP appreciably outperformed the methods DESeq2, ANCOMBC, fitZIG, and MaAsLin2 in inferential accuracy and was reliable even with highly sparse datasets.
Figure 6: For 25% zeros in the abundance matrix, summarizing over the 25 artificial datasets of Section 4, boxplots of AUC (A), FDR (B), and sensitivity (C) for the different statistical methods. The dashed horizontal line in sub-figure B indicates the target FDR of 5% for all the methods.

5 Data analysis

We applied the ZIBNP technique to analyze publicly available microbiome data from the CAMP and Global Gut studies, consisting of $K = 2$ and $K = 3$ groups, respectively. The results were compared with other differential analysis techniques.

CAMP study canine data The dataset was downloaded from the MicrobiomeDB resource [Oliveira et al., 2017]. To investigate the association between eukaryotic parasite infection and the gut microbiome, the investigators processed fecal samples of 155 infected (case) and 115 uninfected (control) dogs and sequenced the V4 region of the 16S rRNA gene. Animal-specific attributes such as sterilization, pet ownership, age, and sex were recorded, allowing statistical methods to adjust for these covariates while detecting the DA taxa.

Table 3 presents the DA taxa detected by ZIBNP between the case and control groups. Some of the detected taxa have been reported by previous studies. For example, bacteria of
|                      | 13%         | 25%         | 50%         | 60%         |
|----------------------|-------------|-------------|-------------|-------------|
| ZIBNP                | 0.79 (0.67,0.95) | 0.77 (0.67,0.96) | 0.68 (0.55,0.83) | 0.58 (0.44,0.69) |
| ANCOMBC              | 0.49 (0.43,0.54) | 0.51 (0.45,0.62) | 0.51 (0.45,0.58) | 0.51 (0.47,0.64) |
| fitZIG               | 0.50 (0.41,0.62) | 0.49 (0.40,0.56) | 0.50 (0.39,0.61) | 0.50 (0.45,0.54) |
| MaAsLin2             | 0.50 (0.41,0.59) | 0.50 (0.46,0.58) | 0.50 (0.43,0.56) | 0.51 (0.47,0.60) |

Table 2: For the simulation study, aggregating over the 25 artificial datasets, average and 95% confidence intervals of the ROC plot AUCs for different DA methods (rows). The columns correspond to different data sparsity levels.

The genus *Bacteroides* were detected as DA with low relative abundance in infected dogs. In studies of the effect of parasite-induced infections on gut microbiome composition, bacteria of the genera *Megamonas* and *Prevotella* were significantly associated with the infection status of dogs [Berry et al., 2020, Pilla and Suchodolski, 2020]. The genus *Blautia* has been reported to have a strong dysbiosis in dogs with acute diarrhea [Pilla and Suchodolski, 2020]. In addition, microbes belonging to the genera *Streptococcus* [Pilla and Suchodolski, 2020], *Ruminococcus gnavus* [Hall et al., 2017, Boucard et al., 2021], and *Alloprevotella* [Boucard et al., 2021] have known associations with inflammatory bowel disease.

The methods ANCOMBC, fitZIG, Maaslin2, DESeq2, and ZIBNP detected 9, 72, 7, 42, and 10 DA taxa, respectively. Table 4 displays the pairwise similarity measures between the different sets of DA taxa detected by the statistical methods using the Jaccard index [Jaccard, 1901]. The Jaccard index \( J \) is a measure of similarity between two sets that ranges between 0% to 100%, and is defined as the size of the intersection set divided by the size of the union set. Whereas \( J = 0\% \) implies that the two sets have no overlap, \( J = 100\% \) implies the two sets are identical. Although all the methods in Table 4 display low overlap, there is moderate overlap between ANCOMBC and Maaslin2, possibly because both methods rely on generalized linear models. The proposed ZIBNP method shares some DA taxa with ANCOMBC and Maaslin2. However, ZIBNP has a low overlap with fitZIG and DESeq2, which respectively rely on zero-inflated Gaussian and negative binomial models.
A Venn diagram of the number of the DA taxa detected by ANCOMBC, MaAsLin2, and ZIBNP is presented in Figure 7. The genera *Blautia* and *Lachnoclostridium* were the common findings of ZIBNP, ANCOMBC, and Maaslin2. Additionally, *Ruminococcus gnavus* was detected by both ZIBNP and ANCOMBC.

### 5.1 Global Gut microbiome study

We applied the proposed ZIBNP method to analyze the motivating Global Gut microbiome data [Yatsunenko et al., 2012]. The study examines the differences in the microbial abundance between samples collected from individuals residing in Malawi, Venezuela, and USA. The dataset consisted of the microbiome abundances of 100 U.S. individuals and 83 individuals each from Malawi and Venezuela, in addition to age and sex as the covariates. The abundance matrix for $n = 266$ subjects and $p = 1,270$ taxa was comprised of 30%
Table 3: In the CAMP study, differentially abundant taxa detected by ZIBNP between infected and uninfected dogs as the groups.
zeros. Along with the proposed ZIBNP approach, we analyzed the data using the methods ANCOMBC, fitZIG and DESeq2. Since MaAsLin2 in its current form does not have a global significance test for more than two groups, we did not include this method in this analysis.

The results are shown in Table 5 and graphically summarized in Figure 8. The four methods detected a common set of 37 DA taxa, as seen in Figure 8. ANCOMBC and fitZIG shared a large proportion of DA taxa and had a Jaccard index of 82%. FitZIG (MetagenomeSeq) declared 1,035 (out of 1,270) taxa as DA. This is consistent with the performance of FitZIG for the CAMP canine dataset, where it also detected the largest number of DA taxa among all the competing methods.

Table S1, available online as part of Supplementary Material, provides the list of 201 DA taxa detected by ZIBNP. Among all DA taxa detected by ZIBNP, 5.5% and 5% of the taxa respectively belonged to the genus *Prevotella* and *Bacteroides*. The original research publication [Yatsunenko et al., 2012] reporting the Global Gut microbiome results focused on pairwise comparisons, such as U.S. versus non-U.S. individuals and Venezuela versus Malawi individuals. Consequently, they are not directly comparable to multigroup DA analysis. However, Yatsunenko et al. [2012] reported several taxa belonging to the genus *Prevotella* as differential for the pairwise geographical regions. Furthermore, they stated that taxa belonging to the genus *Bacteroides* were significantly more abundant in U.S. individuals compared to non-U.S. individuals, in conformity with the findings of the

|            | ANCOMBC | fitZIG | Maaslin2 | DESeq2 | ZIBNP |
|------------|---------|--------|----------|--------|-------|
| ANCOMBC    | 1       | 0.03   | 0.6      | 0.11   | 0.19  |
| fitZIG     | 0.03    | 1      | 0.03     | 0.28   | 0     |
| Maaslin2   | 0.6     | 0.03   | 1        | 0.11   | 0.13  |
| DESeq2     | 0.11    | 0.28   | 0.11     | 1      | 0.02  |
| ZIBNP      | 0.19    | 0      | 0.13     | 0.02   | 1     |

Table 4: For the CAMP study data, pairwise Jaccard index of the DA taxa detected by different statistical methods.
6 Discussion

Differential abundance analyses between multiple groups of study samples help identify novel therapeutic targets for disease treatment. However, few existing methods for differential analysis perform consistently well for different datasets [Weiss et al., 2017]. Motivated by the challenges of high-dimensionality, sparsity, and compositionality typical of microbiome data with covariates, we propose a novel zero-inflated Bayesian nonparametric (ZIBNP) model that adjusts to the distinctive data characteristics. Key contributors to our strategy’s success are (i) the ability to relate the sampling depths (i.e., total counts) to inferential precision while accommodating the inherent compositionality of the data, and (ii) a model-based censoring framework that learns the stochastic relationship between

Figure 8: Venn diagram of the number of DA taxa detected by the methods ANCOMBC, DESeq2, fitZIG (MetagenomeSeq), and ZIBNP in the Global Gut dataset.

proposed ZIBNP approach.
Table 5: For the Global Gut microbiome dataset, pairwise Jaccard index of the DA taxa detected by different statistical methods.

|          | ANCOMBC | fitZIG | DESeq2 | ZIBNP |
|----------|----------|--------|--------|-------|
| ANCOMBC  | 1        | 0.82   | 0.48   | 0.17  |
| fitZIG   | 0.82     | 1      | 0.51   | 0.16  |
| DESeq2   | 0.48     | 0.51   | 1      | 0.06  |
| ZIBNP    | 0.17     | 0.16   | 0.06   | 1     |

the sampling depths and pattern of zeros to effectively impute missing data and, thereby, improve the accuracy of differential analysis.

Latent clusters induced by the nonparametric Chinese restaurant process alleviate collinearity issues due to the small $n$, large $p$ microbiome datasets. The compositional aspects are modeled by a regression framework for the log-ratios of cluster-specific group parameters. The taxa differential statuses are then evaluated as deterministic functions of multivariate group-cluster regression parameters. Through simulation experiments and data analyses, we demonstrate the potential of ZIBNP as a reliable tool of the standard toolbox of wide-ranging microbiome investigations.

As suggested by several studies [e.g., Sankaran and Holmes, 2014, Xiao et al., 2017], the accuracy of statistical methods for differential analysis can be further improved by incorporating the phylogenetic distances between the taxa. Looking ahead, our research will focus on utilizing this valuable information. Commented R code implementing the ZIBNP approach is available on GitHub at https://github.com/archiesach/ZIBNP. We are developing a faster implementation using high-performance Rcpp subroutines that will also be publicly available on GitHub.

References

William G. Wade. The oral microbiome in health and disease. *Pharmacological Research*, 69(1):137–143, 2013. ISSN 1043-6618.
Andrew B. Shreiner, John Y. Kao, and Vincent B. Young. The gut microbiome in health and in disease. *Current Opinion in Gastroenterology*, 31(1):69–75, Jan 2015. 3

S. Rautava. Early microbial contact, the breast milk microbiome and child health. *Journal of Developmental Origins of Health and Disease*, 7(1):5–14, 2016. doi: 10.1017/S2040174415001233. 3

W. H. Wilson Tang, Daniel Y. Li, and Stanley L. Hazen. Dietary metabolism, the gut microbiome, and heart failure. *Nature Reviews Cardiology*, 16(3):137–154, Mar 2019. ISSN 1759-5010. doi: 10.1038/s41569-018-0108-7. URL https://doi.org/10.1038/s41569-018-0108-7. 3

Marco Cappellato, Giacomo Baruzzo, and Barbara Di Camillo. Investigating differential abundance methods in microbiome data: A benchmark study. *PLOS Computational Biology*, 18(9):e1010467, 2022. 3

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, Mar 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0237-y. URL https://doi.org/10.1186/s40168-017-0237-y. 3, 4, 5, 32

Ivan Vujkovic-Cvijin, Jack Sklar, Lingjing Jiang, Loki Natarajan, Rob Knight, and Yasmine Belkaid. Host variables confound gut microbiota studies of human disease. *Nature*, 587(7834):448–454, Nov 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2881-9. URL https://doi.org/10.1038/s41586-020-2881-9. 3, 6

Steven W. Kembel, Martin Wu, Jonathan A. Eisen, and Jessica L. Green. Incorporating 16s gene copy number information improves estimates of microbial diversity and abundance. *PLOS Computational Biology*, 8(10):1–11, 10 2012. doi: 10.1371/journal.pcbi.1002743. URL https://doi.org/10.1371/journal.pcbi.1002743. 3
Kyle Bibby. Metagenomic identification of viral pathogens. *Trends in Biotechnology*, 31(5):275–279, 2013. ISSN 0167-7799. doi: https://doi.org/10.1016/j.tibtech.2013.01.016. URL https://www.sciencedirect.com/science/article/pii/S0167779913000292.

David A. Relman. Metagenomics, Infectious Disease Diagnostics, and Outbreak Investigations: Sequence First, Ask Questions Later? *JAMA*, 309(14):1531–1532, 04 2013. ISSN 0098-7484. doi: 10.1001/jama.2013.3678. URL https://doi.org/10.1001/jama.2013.3678.

Thomas J. Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 2014. ISSN 1664-462X. doi: 10.3389/fpls.2014.00209. URL https://www.frontiersin.org/article/10.3389/fpls.2014.00209.

Richa Bharti and Dominik G Grimm. Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 22(1):178–193, 12 2019. ISSN 1477-4054. doi: 10.1093/bib/bbz155. URL https://doi.org/10.1093/bib/bbz155.

Yinglin Xia, Jun Sun, and Ding-Geng Chen. *Bioinformatic Analysis of Microbiome Data*, pages 1–27. Springer Singapore, Singapore, 2018a. ISBN 978-981-13-1534-3. doi: 10.1007/978-981-13-1534-3_1. URL https://doi.org/10.1007/978-981-13-1534-3_1.

Samuel Anyaso-Samuel, Archie Sachdeva, Subharup Guha, and Somnath Datta. Metagenomic geolocation prediction using an adaptive ensemble classifier. *Frontiers in Genetics*, 12:642282, 2021.

Patrick D Schloss and Sarah L Westcott. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rRNA gene sequence analysis. *Applied and Environmental Microbiology*, 77(10):3219–3226, 2011.

J. Aitchison. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. *Proceedings of CoDaWork ’08, The 3rd Compositional Data Analysis Workshop, Girona, Spain*, 2008.
Donald T. McKnight, Roger Huerlimann, Deborah S. Bower, Lin Schwarzkopf, Ross A. Alford, and Kyall R. Zenger. Methods for normalizing microbiome data: An ecological perspective. *Methods in Ecology and Evolution*, 10(3):389–400, 2019. doi: https://doi.org/10.1111/2041-210X.13115. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13115.

María Bailén, Carlo Bressa, Sara Martínez-López, Rocío Gonzále-Soltero, Maria Gregoria Montalvo Lominchar, Celia San Juan, and Mar Larrosa. Microbiota features associated with a high-fat/low-fiber diet in healthy adults. *Frontiers in Nutrition*, 7:583608, 2020.

Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2020.09.014. URL https://www.sciencedirect.com/science/article/pii/S2001037020303986.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology*, 8:2114–2114, Nov 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02114. URL https://pubmed.ncbi.nlm.nih.gov/29163406. 29163406[pmid].

Zachary D. Wallen. Comparison study of differential abundance testing methods using two large parkinson disease gut microbiome datasets derived from 16s ampli-con sequencing. *BMC Bioinformatics*, 22(1):265, May 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04193-6. URL https://doi.org/10.1186/s12859-021-04193-6.

Somnath Datta, Subharup Guha, et al. *Statistical Analysis of Microbiome Data*. Springer, 2021.

Peng Liu, Emily Goren, Paul Morris, David Walker, and Chong Wang. *Statistical Methods for Feature Identification in Microbiome Studies*, pages 175–192. Springer International
Jacob T. Nearing, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau, and Morgan G. I. Langille. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):342, Jan 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28034-z. URL https://doi.org/10.1038/s41467-022-28034-z

Fernandes, A. D., Reid, J., Macklaim, J. M., McMurrough, T.A, Edgell, D.R., Gloor, and G. B. Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. ***Microbiome, 2014, volume 2, 15, 2014. URL ***http://doi:10.1186/2049-2618-2-15

Joseph N. Paulson, O. Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, Dec 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2658. URL https://doi.org/10.1038/nmeth.2658

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1):284, Jan 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02554-5. URL https://doi.org/10.1038/s41467-017-02554-5

Yinglin Xia, Jun Sun, and Ding-Geng Chen. *Modeling Zero-Inflated Microbiome Data*, pages 453–496. Springer Singapore, Singapore, 2018b. ISBN 978-981-13-1534-3. doi: 10.1007/978-981-13-1534-3_12. URL https://doi.org/10.1007/978-981-13-1534-3_12

Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):3514, Jul 2020. ISSN 2041-1723. doi:
Himel Mallick, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, George Weingart, Boyu Ren, Emma H. Schwager, Suvo Chatterjee, Kelsey N. Thompson, Jeremy E. Wilkinson, Ayshwarya Subramanian, Yiren Lu, Levi Waldron, Joseph N. Paulson, Eric A. Franzosa, Hector Corrada Bravo, and Curtis Huttenhower. Multivariable association discovery in population-scale meta-omics studies. *PLOS Computational Biology*, 17(11):1–27, 11 2021. doi: 10.1371/journal.pcbi.1009442. URL https://doi.org/10.1371/journal.pcbi.1009442.

Michael I Love, Wolfgang Huber, and Simon Anders. *Genome Biology*, 15(12):1–21, 2014.

Bryan D Martin, Daniela Witten, and Amy D Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1):94, 2020.

Peter Müller and Riten Mitra. Bayesian Nonparametric Inference – Why and How. *Bayesian Analysis*, 8(2):269 – 302, 2013. doi: 10.1214/13-BA811. URL https://doi.org/10.1214/13-BA811.

A. Lijoi and I. Prünster. *Models beyond the Dirichlet process*, pages 80–136. Cambridge Series in Statistical and Probabilistic Mathematics, 2010.

Francislon S Oliveira, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, Jessica C Kissinger, Brian P Brunk, Jr Stoeckert, Christian J, Gabriel R Fernandes, David S Roos, and Daniel P Beiting. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Research*, 46(D1):D684–D691, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1027. URL https://doi.org/10.1093/nar/gkx1027.

Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P
Anokhin, et al. Human gut microbiome viewed across age and geography. *Nature*, 486 (7402):222–227, 2012. 5, 29, 31

John Lee, Manthan Shah, Sara Ballouz, Megan Crow, and Jesse Gillis. Cococonet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Research*, 48(W1):W566–W571, 2020. 8

M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20:1222–1232, 2004. 8

S. Kim, M. G. Tadesse, and M. Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93:877–893, 2006. 8

D. B. Dunson, A. H. Herring, and S. M. Engel. Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association*, 103:534–546, 2008. 8

D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008. 8

Subharup Guha, Rex Jung, and David Dunson. Predicting phenotypes from brain connection structure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71 (4):639–668, 2022. 8

Chiyu Gu, Veerabhadran Baladandayuthapani, and Subharup Guha. Nonparametric bayes differential analysis of multigroup dna methylation data. *Bayesian Analysis*, 2023. To appear. 8

Yuanjing Ma, Yuan Luo, and Hongmei Jiang. A novel normalization and differential abundance test framework for microbiome data. *Bioinformatics*, 36(13):3959–3965, 04 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa255. URL https://doi.org/10.1093/bioinformatics/btaa255. 14
W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992. ISSN 00359254, 14679876. URL [http://www.jstor.org/stable/2347565](http://www.jstor.org/stable/2347565). 16

M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, Apr 2004. 19

Zachary M. Burcham, et al., and Genetics Taste Lab Citizen Scientists. Patterns of oral microbiota diversity in adults and children: A crowdsourced population study. *Scientific Reports*, 10(1):2133, Feb 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-59016-0. URL [https://doi.org/10.1038/s41598-020-59016-0](https://doi.org/10.1038/s41598-020-59016-0). 19

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: [https://doi.org/10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x). URL [https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x](https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x). 24

Jonathan Thorsen, Asker Brejnrod, Martin Mortensen, Morten A Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, 4:1–14, 2016. 26

Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings*
Alexander SF Berry, Kaylynn Johnson, Rene Martins, Megan C Sullivan, Camila Farias Amorim, Alexandra Putre, Aiysha Scott, Shuai Wang, Brianna Lindsay, Robert N Baldassano, et al. Natural infection with giardia is associated with altered community structure of the human and canine gut microbiome. *Msphere*, 5(4):10–1128, 2020. 28

Rachel Pilla and Jan S. Suchodolski. The role of the canine gut microbiome and metabolome in health and gastrointestinal disease. *Frontiers in Veterinary Science*, 6, 2020. ISSN 2297-1769. doi: 10.3389/fvets.2019.00498. URL https://www.frontiersin.org/article/10.3389/fvets.2019.00498. 28

Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, Madeline Bertha, Melissa Cohen, John Garber, Hamed Khalili, Dirk Gevers, Ashwin Ananthakrishnan, Subra Kugathasan, Eric Lander, Paul Blainey, and Curtis Huttenhower. A novel ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine*, 9, 11 2017. doi: 10.1186/s13073-017-0490-5. 28

Anne-Sophie Boucard, Myriam Thomas, Wilfried Lebon, Bruno Polack, Isabelle Florent, Philippe Langella, and Luis G. Bermúdez-Humarán. Age and giardia intestinalis infection impact canine gut microbiota. *Microorganisms*, 9(9):1862, Sep 2021. ISSN 2076-2607. doi: 10.3390/microorganisms9091862. URL https://pubmed.ncbi.nlm.nih.gov/34576757. 28

Paul Jaccard. Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 01 1901. doi: 10.5169/seals-266450. 28

Kris Sankaran and Susan Holmes. structssi: simultaneous and selective inference for grouped or hierarchically structured data. *Journal of Statistical Software*, 59(13):1–21, 2014. 33
Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18):2873–2881, 05 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx311. URL https://doi.org/10.1093/bioinformatics/btx311.