METHODS FOR SINGLE-CELL AND MICROBIOME SEQUENCING DATA

EDITED BY: Himel Mallick, Lingling An, Mengjie Chen, Pei Wang and Ni Zhao

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-280-4
DOI 10.3389/978-2-88976-280-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

METHODS FOR SINGLE-CELL AND MICROBIOME SEQUENCING DATA

Topic Editors:
Himel Mallick, Merck, United States
Lingling An, University of Arizona, United States
Mengjie Chen, The University of Chicago, United States
Pei Wang, Icahn School of Medicine at Mount Sinai, United States
Ni Zhao, Johns Hopkins University, United States

Citation: Mallick, H., An, L., Chen, M., Wang, P., Zhao, N., eds. (2022). Methods for Single-Cell and Microbiome Sequencing Data. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-280-4

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 13 May 2022
doi: 10.3389/fgene.2022.920191

Editorial: Methods for Single-Cell and Microbiome Sequencing Data

Himel Mallick[1]*, *Lingling An*[2,3,4], *Mengjie Chen*[5], *Pei Wang*[6,7] and *Ni Zhao*[8]

[1]Biostatistics and Research Decision Sciences, Merck & Co.Inc., Rahway, NJ, United States, [2]Interdisciplinary Program in Statistics and Data Science, The University of Arizona, Tucson, AZ, United States, [3]Department of Epidemiology and Biostatistics, The University of Arizona, Tucson, AZ, United States, [4]Department of Biosystems Engineering, The University of Arizona, Tucson, AZ, United States, [5]Department of Human Genetics and Department of Medicine, University of Chicago, Chicago, IL, United States, [6]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [7]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [8]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, United States

Keywords: microbiome, single-cell, omics, data science, multi-omics, statistics, biostatistics, computational biology

Editorial on the Research Topic

Methods for Single-Cell and Microbiome Sequencing Data

Translational investigations of single-cell transcriptomics and microbiomics now constitute the research hotspots in the field of omics sciences with cell-type-specific gene expression and host-associated microbes and microbial gene products implicated in numerous complex diseases (Mallick et al., 2017; Aldridge and Teichmann, 2020). Motivated by the structural similarities of scRNAseq and metagenomics data (Calgaro et al., 2020; Jeganathan and Holmes, 2021), with respect to several statistical properties such as, high-dimensionality, count and compositional nature, excess zeros due to low sequencing depth or dropout, overdispersion, and spatial and temporal dependence, among others, we set out to launch a combined Research Topic following the completion of the successful first volume (Mallick et al., 2020) in 2020.

This Research Topic thus consists of eleven papers (including the editorial) on various since-cell and microbiome omics areas and covers the latest development of statistical methods for analyzing microbiome and single-cell sequencing data. The papers can be broadly categorized into four subtypes (**Figure 1**): 1) Specialized domain-specific publications, 2) domain-agnostic publications applicable to both microbiome and single-cell studies, 3) single-cell-specific methods with potential applicability to microbiome studies, and 4) microbiome-specific methods with potential applicability to scRNASeq.

One of the most common applications of omics data is the differential expression or abundance analysis to identify omics features that are differential between two or more biological conditions. Despite being a well-studied problem, differential analysis is still a very active area of research. In both single-cell and microbiome studies, given the large number of features present in a typical dataset, standard statistical testing procedures can put false association or loss of power at odds with prior knowledge or expectations (Mallick et al., 2017). While most of the current methods are domain- or platform-specific, domain-agnostic methods applicable to multiple platforms or data types are becoming increasingly common (Mallick et al., 2021a; Rahnavard et al., 2021). Taking advantage of the inherent compositionality and hierarchical tree structure observed in both single-cell and microbiome sequencing data, Ostner et al. proposes a domain-agnostic Bayesian tree-aggregated model (tascCODA) applicable to any compositional rectangular data with hierarchical row or column information. tascCODA thus constitutes a valuable addition to the growing statistical

Edited and reviewed by:
Richard D. Emes, University of Nottingham, United Kingdom

***Correspondence:**
Himel Mallick
himel.mallick@merck.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *14 April 2022*
Accepted: *26 April 2022*
Published: *13 May 2022*

Citation:
Mallick H, An L, Chen M, Wang P and Zhao N (2022) Editorial: Methods for Single-Cell and Microbiome Sequencing Data. Front. Genet. 13:920191. doi: 10.3389/fgene.2022.920191

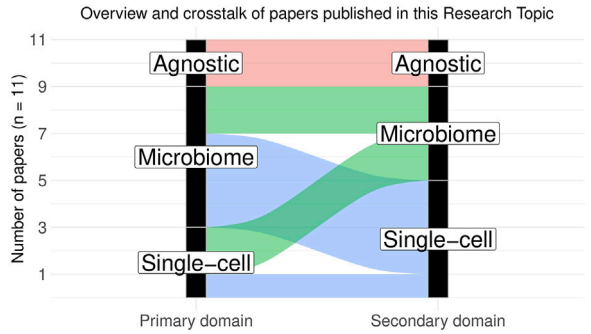


FIGURE 1 | Overview and crosstalk of papers published in this Research Topic. The papers can be broadly categorized into four subtypes: 1) domain-agnostic publications generally applicable to both microbiome and single-cell studies (2 papers), 2) single-cell-specific methods that can be potentially applied to microbiome data with necessary modifications (2 papers), 3) microbiome-specific methods that can be potentially applied to scRNASeq data with necessary modifications (4 papers) and 4) domain-specific analysis methods or literature reviews (1 single-cell-specific and 2 microbiome-specific papers). Primary domain indicates the intended application area for the related paper, whereas secondary domain indicates the potential application area wherein the said method can be extended with necessary modifications.

toolbox of domain-agnostic methods for omics research enhancing interoperability of disparate omics datasets (Sansone et al., 2009; Conesa and Beck, 2019).

A popular alternative to per-feature differential abundance analysis methods is the community-level or omnibus association methods that enable associating the entire microbial community composition with a phenotype of interest (Mallick et al., 2017). Due to their multivariate setups, omnibus association methods typically fail to provide feature-level inference to enable follow-up characterization (Mallick et al., 2021b). To this end, Chen et al. proposes a hybrid method (MiAF) that adaptively combines *p*-values from the feature-level tests to construct a community-level test, thus providing the best of both worlds in a unified framework. Jiang et al. extends the popular community-level test (MiRKAT) to multi-categorical nominal and ordinal outcomes for both independent or clustered (e.g., family-based and longitudinal) microbiome studies.

Keeping pace with ongoing advances in artificial intelligence, a variety of machine learning methods have become available to analyze microbiome and single-cell data. Deek and Li proposes a Bayesian data generative process for microbiome community data by developing a zero-inflated Latent Dirichlet Allocation (zinLDA) model that accurately identifies the latent sparse subcommunities of a microbial community, improving upon the state-of-the-art Latent Dirichlet Allocation (LDA) model. Zhang et al. develops a novel, unsupervised, data-driven deep learning-based imputation method (NISC) to impute the excess amount of zeroes (dropouts) observed in scRNA-seq count data that improves downstream cell type identification accuracy compared to existing imputation methods.

Just as differential analysis provides one potential area to transfer methods between fields, inference of feature-feature interaction network estimation provides another. Improving upon the existing cross-sectional ecological network inference methods, He et al. proposes a novel autoregressive zero-inflated Poisson mixed-effects model (ARZIMM) to detect sparse microbial interactions in longitudinal microbiome data, thus providing a scalable alternative to existing computationally intensive temporal ecological network detection and stability estimation methods.

Both microbial community and single-cell datasets possess unique characteristics that differ in ways that necessitate the development of domain-specific tools, with many of the single-omics tools not susceptible to technological variability induced by experimental platforms or library preparation protocols (Mallick et al., 2021a). To this end, several domain- and platform-specific methods and literature reviews have been published to better address the biological question at hand within a specific context.

Wu et al. proposes a non-linear normalization approach for non-UMI single-cell data that reduces more technical variation than competing methods without reducing biological variation. Jones et al. asserts that in 16S rRNA gene sequencing data (specially in the Ion Torrent platform), assessing multiple hypervariable regions in tandem is critical to enhance the statistical evaluation of overall differences in community structure and relatedness among samples. Paisley and Liu develops and deploys an R Shiny web tool (GeneMarkeR) in order to provide a vastly expanded, standardized marker gene database for the end users, improving upon existing overwhelmingly incoherent databases often with a lack of validated standards. Finally, Arbas et al. carefully curates the literature to highlight the current state-of-the-field in longitudinal microbiome studies ranging from experimental design and basic bioinformatics preprocessing steps to critical multi-omic data integration considerations including modeling, validation, and inference.

Many of the methods described in this Research Topic also come with accompanying open-source software implementations, thus providing an important resource for future methodologists and machine learners and many of them are potentially extensible to other data types beyond their intended application domains (**Figure 1**). As the field of omics research progresses, we expect to see more research linking disparate omics data with human genetics and digital pathology in order to gain better functional insights into the role of omics features in disease initiation and progression. We also expect to see more diverse data sets at the intersection of spatial omics, long-read sequencing, and imaging genomics, giving rise to new statistical questions and challenges, which motivated us to launch a third volume of the Research Topic on imaging and omics data science. We hope that omics and imaging scientists from various subfields will work together in this exciting area of research and make important scientific contributions by providing a shared infrastructure for common data types and fostering ideas for more sophisticated, reproducible, interpretable data analyses.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We thank the Frontiers editorial staff for providing outstanding assistance in putting together this Research Topic collection.

REFERENCES

Aldridge, S., and Teichmann, S. A. (2020). Single Cell Transcriptomics Comes of Age. *Nat. Commun.* 11, 4307. doi:10.1038/s41467-020-18158-5

Calgaro, M., Romualdi, C., Waldron, L., Risso, D., and Vitulo, N. (2020). Assessment of Statistical Methods from Single Cell, Bulk RNA-Seq, and Metagenomics Applied to Microbiome Data. *Genome Biol.* 21, 191. doi:10.1186/s13059-020-02104-1

Conesa, A., and Beck, S. (2019). Making Multi-Omics Data Accessible to Researchers. *Sci. Data* 6, 251. doi:10.1038/s41597-019-0258-4

Jeganathan, P., and Holmes, S. P. (2021). A Statistical Perspective on the Challenges in Molecular Microbial Biology. *J. Agric. Biol. Environ. Statistics* 26, 131–160. doi:10.1007/s13253-021-00447-1

Mallick, H., Ma, S., Franzosa, E. A., Vatanen, T., Morgan, X. C., and Huttenhower, C. (2017). Experimental Design and Quantitative Analysis of Microbial Community Multiomics. *Genome Biol.* 18, 228. doi:10.1186/s13059-017-1359-z

Mallick, H., Bucci, V., and An, L. (2020). Editorial: Statistical and Computational Methods for Microbiome Multi-Omics Data. *Front. Genet.* 11, 927. doi:10.3389/fgene.2020.00927

Mallick, H., Chatterjee, S., Chowdhury, S., Chatterjee, S., Rahnavard, A., and Hicks, S. C. (2021a). Differential Expression of Single-Cell RNA-Seq Data Using Tweedie Models. *bioRxiv*. doi:10.1101/2021.03.28.437378

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021b). Multivariable Association Discovery in Population-Scale Meta-Omics Studies. *PLoS Comput. Biol.* 17, e1009442. doi:10.1371/journal.pcbi.1009442

Rahnavard, A., Chatterjee, S., Sayoldin, B., Crandall, K. A., Tekola-Ayele, F., and Mallick, H. (2021). Omics Community Detection Using Multi-Resolution Clustering. *Bioinformatics* 37, 3588–3594. doi:10.1093/bioinformatics/btab317

Sansone, S. A., Rocca-Serra, P., Field, D., Taylor, C. F., Tong, W., Brandizi, M., et al. (2009). Towards Interoperable Reporting Standards for Omics Data: Hopes and Hurdles. *Summit Transl. Bioinform* 2009, 112–115.

Conflict of Interest: HM is employed by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 22 January 2021
doi: 10.3389/fgene.2020.602594

A Zero-Inflated Latent Dirichlet Allocation Model for Microbiome Studies

*Rebecca A. Deek and Hongzhe Li**

Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

The human microbiome consists of a community of microbes in varying abundances and is shown to be associated with many diseases. An important first step in many microbiome studies is to identify possible distinct microbial communities in a given data set and to identify the important bacterial taxa that characterize these communities. The data from typical microbiome studies are high dimensional count data with excessive zeros due to both absence of species (structural zeros) and low sequencing depth or dropout. Although methods have been developed for identifying the microbial communities based on mixture models of counts, these methods do not account for excessive zeros observed in the data and do not differentiate structural from sampling zeros. In this paper, we introduce a zero-inflated Latent Dirichlet Allocation model (zinLDA) for sparse count data observed in microbiome studies. zinLDA builds on the flexible Latent Dirichlet Allocation model and allows for zero inflation in observed counts. We develop an efficient Markov chain Monte Carlo (MCMC) sampling procedure to fit the model. Results from our simulations show zinLDA provides better fits to the data and is able to separate structural zeros from sampling zeros. We apply zinLDA to the data set from the American Gut Project and identify microbial communities characterized by different bacterial genera.

Keywords: metagenomics, gibbs sampling, zero inflated dirchlet distribution, mixture models, microbial community

OPEN ACCESS

Edited by:
Lingling An,
University of Arizona, United States

Reviewed by:
Huilin Li,
New York University, United States
Koichi Higashi,
National Institute of Genetics, Japan

***Correspondence:**
Hongzhe Li
hongzhe@upenn.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *03 September 2020*
Accepted: *29 December 2020*
Published: *22 January 2021*

Citation:
Deek RA and Li H (2021) A Zero-Inflated Latent Dirichlet Allocation Model for Microbiome Studies. Front. Genet. 11:602594. doi: 10.3389/fgene.2020.602594

1. INTRODUCTION

The advent and proliferation of next-generation sequencing (NGS) technologies has given rise to many large-scale high-throughput microbiome studies (Turnbaugh et al., 2007; Gilbert et al., 2014; McDonald et al., 2018). Classical statistical techniques are not able to evaluate such data due to its inherent high dimensional, count-based, and sparse nature. Consequently, novel statistical methods are necessary for accurate and unbiased analysis of such data.

Much of microbiome research has focused on high-dimensional statistical methods, as a single 16S rRNA gene sequencing sample can produce tens of thousands of sequencing reads from hundreds of different amplicon sequence variants (ASVs). Of particular interest are techniques for dimensionality reduction. Commonly used methods include principal coordinate analysis (PCoA) with distance measures, such as weight and unweighted UniFrac distance and Bray-Curtis dissimilarity, or canonical correlation analysis with sparsity assumptions (Chen et al., 2013; Hawinkel et al., 2019). More recently, studies have begun to focus on understanding microbial

dynamics within the human microbiome. Single-species analysis, that focus on one species at a time in a "parts-list" fashion, are not able to capture complex and dynamic interactions. These inter-species interactions form the basis of distinct underlying subcommunity structures and failing to account for them contributions to the data heterogeneity commonly seen in microbiome studies. As such, network-based approaches have been successfully applied in this area (Faust and Raes, 2012; Layeghifard et al., 2017). These methods use co-occurrence or correlation measures to identify pairwise interactions in cross-sectional studies (Faust et al., 2012; Friedman and Alm, 2012; Kurtz et al., 2015). Others use temporally conserved covariance to identify interactions in longitudinal studies (Raman et al., 2019).

Generative probabilistic mixture models are able to act as a dimensionality reduction technique while simultaneously describing microbial dynamics via subcommunity identification. When applied to microbiome data the latent variable(s) in a mixture model have meaningful biological connotations. Specifically, they represent distinct subcommunity profiles, or structures, that give rise to the observed samples. The simplest of these is the Dirichlet-multinomial mixture model (Holmes et al., 2012). This model is a generalization of the Dirichlet-multinomial hierarchical model. Rather than assuming that all samples in a cohort are generated from a single community profile, as the Dirichlet-multinomial model does, the mixture model assumes the cohort contains many different subcommunity structures and each of the samples is generated by one of them (Holmes et al., 2012). As such, a sample can be described by its subcommunity assignment rather than a high-dimensional vector of ASV counts. Though, the Dirichlet-multinomial mixture model may still be too restrictive to accurately capture microbial community structures and all the heterogeneity of microbiome studies (Sankaran and Holmes, 2019). It is biologically plausible that an individual's microbiome is comprised of numerous subcommunities, rather than just one, mixing together to varying degrees. The Latent Dirichlet Allocation (LDA) model describes such a generative process (Blei et al., 2003). Samples are defined by their mixture probabilities for each of the subcommunities rather than belonging to a single one. Technically speaking, LDA differs from the Dirichlet-multinomial mixture model by sampling the latent community variable repeatedly within a sample, once per sequencing read, rather than just once for the entire sample (Blei et al., 2003; Griffiths and Steyvers, 2004).

Latent Dirichlet Allocation has been successful in identifying functional subcommunities of the human gut and skin microbiota (Higashi et al., 2018; Sankaran and Holmes, 2019; Hosoda et al., 2020; Sommeria-Klein et al., 2020). Despite this, it has been noted that LDA is prone to over-smoothing of microbial counts, which are known to be sparse (Sankaran and Holmes, 2019). This can be attributed to the Dirichlet distribution being insufficient to capture the over-dispersion and zero-inflation of microbiome data. The distribution only has one dispersion parameter and inherently imposes a negative correlation between component counts, which may lead to spurious associations (Tang and Chen, 2019). Moreover, the model assumes that each species has a non-negative probability of belonging to every subcommunity. This implies that all species contribute to every subcommunity, even if only with low probability. Although, it is more likely that the presence of one species in a community prevents the presence of another.

As such, it would be advantageous to be able to identify community structures that are only composed of a subset of microbial species present in a data set. Thus, estimating some of the taxa membership probabilities for each subcommunity to be zero. We propose a zero-inflated Latent Dirichlet Allocation (zinLDA) model that is flexible enough to capture sparse subcommunities of microbiota. In the following section we detail the generative process of the LDA model and our zero-inflated LDA model. We also provide information on how to estimate model parameters using Markov chain Monte Carlo (MCMC) methods. We apply both models to simulation studies and real data analysis using data from the American Gut Project to directly compare the two and highlight how our proposed method provides better fit to microbiome data.

2. MATERIALS AND METHODS

2.1. Notation and Terminology

Data in microbiome studies often comes from high-throughput sequencing of the 16S rRNA gene. A single biological sample can be represented by a vector of taxon counts with each component representing the number of reads aligned to that specific classification (e.g., ASV, species, genus). The following definitions and notations will be of help in defining a generative probabilistic model for microbiome studies:

- w_{dn} is the *nth* observed sequencing read in the *dth* biological sample. Sequencing reads are represented by V-length vectors with a single non-zero component whose value is equal to one, where V is the number of unique taxa in the study.
- w_{dn}^{i} represents that the *nth* sequencing read in the *dth* sample belongs to the *ith* unique taxa ($i = 1, \ldots, V$).
- $\boldsymbol{w}_d = (w_{d1}, \ldots, w_{dN})$ is the *dth* biological sample consisting of N sequencing reads.
- A cohort $\mathbf{D} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D)$ is a collection of all biological samples in the study.

2.2. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a probabilistic model that is flexible enough to describe the generative process for discrete data in a variety of fields from text analysis to bioinformatics. When applied to microbiome studies, LDA provides the following generative process for the taxon counts in a cohort **D**:

1. For each of the K subcommunities, indexed by j:
   a. Choose $\boldsymbol{\beta}^{(j)} \sim \text{Dir}(\eta)$
2. For each biological sample $\boldsymbol{w_d}$ in the cohort:
   a. Choose $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha)$
3. For each of the N sequencing reads, w_{dn}:
   a. Choose a subcommunity, $z_{dn} \sim \text{Multinomial}(1, \boldsymbol{\theta}^{(d)})$

b. Choose a taxon w_{dn} from $P(w_{dn}|z_{dn}, \boldsymbol{\beta})$, a multinomial probability distribution conditional on the subcommunity z_{dn}.

Figure 1 provides a graphical model representation of LDA. In this model, $\boldsymbol{\beta} = [\beta_{ij}]$ fully describes the taxa distribution for each subcommunity. The probability that the *ith* taxa belongs to the *jth* subcommunity is denoted by β_{ij}. Note that the taxa distribution is cohort-specific meaning that it is common across all samples and is only estimated once per cohort. The mixture probabilities for the subcommunities of the *dth* sample are denoted by a K-length vector, $\boldsymbol{\theta}^{(d)}$, with θ_{dj} representing the mixture probability of the *jth* subcommunity in the *dth* sample. Here, K is the number of underlying subcommunities and is assumed to be known a-priori. Additionally, z_{dn} is the subcommunity assignment for sequencing read w_{dn}. Both hyperparameters η and α are assumed to be symmetric and are defined once for the whole cohort.

Intuitively, $\beta_{ij} = P(w^i_{dn}|z_{dn} = j)$ determines which taxa are important to subcommunity j and $\theta_{dj} = P(z_{dn} = j)$ determines which subcommunities are important in the *dth* sample. Moreover, the LDA model acts as a "soft" clustering technique by allowing samples to be composed of multiple subcommunities. Geometrically, the parameter space of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be thought of in terms of a simplex space. The taxa per subcommunity distribution belongs the V-1 simplex, such that $\boldsymbol{\beta}^{(j)} \in S^{V-1}$. Meanwhile, $\boldsymbol{\theta}^{(d)}$, the subcommunity distribution per sample can be represented by a randomly selected point in the (K-1)-dimensional simplex, S^{K-1}. This is different from the Dirichlet-Multinomial mixture model in which $\boldsymbol{\theta}^{(d)} = \boldsymbol{\theta}$ is assumed to be fixed across all samples and can be represented by the vertices of S^{K-1}.

2.3. Zero-Inflated Latent Dirichlet Allocation (zinLDA)

We propose a modification to the Latent Dirichlet Allocation model that allows the latent subcommunity organization to be composed of both structural zeros, taxa that truly do not belong to the community, and sampling zeros, taxa that belong to the community, but are not captured due to low sequencing depth or dropout. Understanding and identifying the structural zeros in the data is biologically interesting as it provides insights into the absence of certain taxa in a given community.

The zero-inflated generalized Dirichlet (ZIGD) distribution is able to model both sources of zeros. The generalized Dirichlet (GD) distribution is an extension of the Dirichlet that allows for a more flexible covariance structure via the introduction of additional parameters (Connor et al., 1969). Though, it should be noted that the GD distribution alone does not model structural zeros.

To do so, we must modify the unique relationship between the GD distribution and a set of mutually independent beta random variables. By adding a zero-inflation probability, π, to each of the beta random variables we arrive at the zero-inflated generalized Dirichlet distribution. Formally, a length-V vector of ZIGD compositions, denoted by $\boldsymbol{\beta} = \{\beta_1, \ldots \beta_V\}$, can be formulated from a set of mutually independent zero-inflated beta random variables, which we denote by $\mathbf{Q} = \{Q_1, \ldots, Q_{V-1}\}$, with zero-inflation probabilities, $\boldsymbol{\pi} = \{\pi_1, \ldots \pi_{V-1}\}$ and the parameters in the beta distributions denoted by (a, b). The relationship between the two random variables can be described as follows: $\beta_1 = Q_1$, $\beta_i = \prod_{l=1}^{i-1}(1 - Q_l)$ for $l = 2, \ldots V - 1$, and $\beta_V = \sum_{i=1}^{V-1} \beta_i$ (Tang and Chen, 2019). Furthermore, we introduce an indicator variable, $\Delta_i = I(\beta_i = 0) = I(Q_i = 0)$, to identify structural zeros.

For every subcommunity j, let there be L_j taxa with $\beta_{ij} > 0 \Leftrightarrow \Delta_{ij} = 0$. Then let $\mathbf{U}_j$ denote the set of indices of the non-zero taxa probabilities for subcommunity j, $\mathbf{U}_j = \{u_{1_j}, ..., u_{L_j}\}$, and $\bar{\boldsymbol{U}}_j$ be its complement.

Replacing the Dirichlet(η) prior on $\boldsymbol{\beta}$ with a ZIGD(π, a, b) gives a zero-inflated Latent Dirichlet Allocation (zinLDA) model. The zinLDA model assumes the following generative process for a cohort **D**:

1. For each of the K subcommunities, indexed by j:
   a. Choose $\boldsymbol{\Delta}^{(j)} \sim \text{Ber}(\pi)$
   b. Choose $\boldsymbol{\beta}^{(j)} \sim \text{ZIGD}(\pi, a, b)$
2. For each biological sample $\boldsymbol{w_d}$ in the cohort:
   a. Choose $\boldsymbol{\theta}^{(\boldsymbol{d})} \sim \text{Dir}(\alpha)$
3. For each of the N sequencing reads, w_{dn}:
   a. Choose a subcommunity, $z_{dn} \sim \text{Multinomial}(1, \boldsymbol{\theta}^{(d)})$
   b. Choose a taxon, w_{dn} from $P(w_{dn}|z_{dn}, \boldsymbol{\beta})$, a multinomial probability distribution conditional on the subcommunity z_{dn}.

In this model we assume hyperparameters π, a, b, and α are symmetric and are defined once for the whole cohort. Comparing the graphical model representation of zinLDA to that of the LDA model (**Figure 1**) underscores the differences between the two, particularly with respect to modeling $\boldsymbol{\beta}$.

We adopt a Bayesian framework for inference and parameter estimation. As such, inference for the zinLDA model is centered around the posterior distribution:

$$P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta}|\mathbf{w}; \alpha, \pi, a, b) = \frac{P(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta}, \mathbf{w}|\alpha, \pi, a, b)}{P(\mathbf{w}|\alpha, \pi, a, b)}. \quad (1)$$

Calculation of this distribution cannot be done directly because the marginalization required to find the normalizing constant, $P(\mathbf{w}|\alpha, \pi, a, b)$, is intractable. As such, approximate methods are necessary for parameter estimation. Variational inference may be used to find parameter estimates by maximizing an approximation to the true posterior. Alternatively, a Markov chain Monte Carlo procedure, such as Gibbs sampling, may be used to generate samples from the target posterior distribution for inference. It is worthy to note that due to the fact that both the Dirichlet and ZIGD distributions are conjugate prior for the multinomial distribution using a collapsed Gibbs sampler, marginalizing over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, gives a tractable solution, even more so than had collapsing not been performed. For this reason,

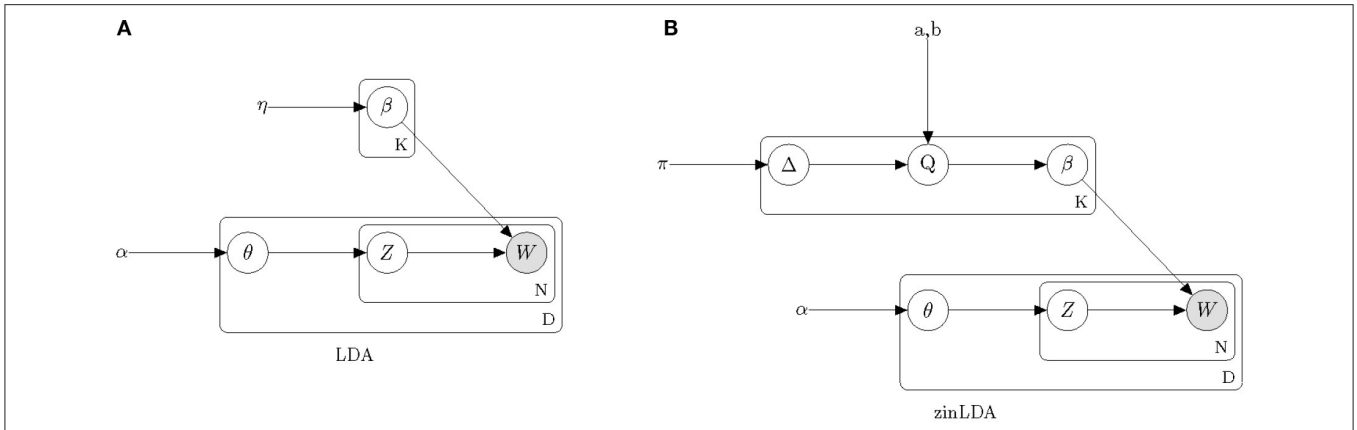


FIGURE 1 | Plate diagrams of Latent Dirichlet Allocation **(A)** and zero-inflated Latent Dirichlet Allocation **(B)**. Nodes represent parameters and random variables, shading denotes the observed data. Boxes represent repeated sampling. The outer box is denoted with D for once per biological sample, the inner box with N for per sequencing read and the upper box with K for per subcommunity.

we proposed a collapsed Gibbs sampler for the joint posterior distribution of $\mathbf{z}$ and $\boldsymbol{\Delta}$ over taxa, $P(\mathbf{z}, \boldsymbol{\Delta}|\mathbf{w})$, where:

$$P(\mathbf{z}, \boldsymbol{\Delta}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Delta})}{P(\mathbf{w})} = \frac{P(\mathbf{w}|\mathbf{z}, \boldsymbol{\Delta})P(\mathbf{z})P(\boldsymbol{\Delta}|\pi)}{\sum_{\Delta}\sum_{z} P(\mathbf{w}, \mathbf{z}, \boldsymbol{\Delta})} \tag{2}$$

Integration over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be done separately as the former only appears in $P(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Delta})$ and the latter only in $P(\mathbf{z}|\boldsymbol{\theta})$. In Gibbs sampling, each state of the chain is taken as an assignment of each z_{dn} and Δ_{ij}. These states are sampled conditional on the observed data and all the other parameters in the model at their current state. Thus, to perform the sampling, the full conditional distributions, $P(z_{dn} = j|\mathbf{w}, \boldsymbol{z}_{-n}, \boldsymbol{\Delta})$ and $P(\Delta_{ij} = 1|\mathbf{w}, \mathbf{z}, \boldsymbol{\Delta}_{-i})$, must be known. These distributions have closed form solutions due to the conjugate prior property of the Dirichlet and ZIGD distributions and can be found probabilistically (**Supplementary Material**):

$$P(z_{dn} = j|\boldsymbol{z}_{-n}, \mathbf{w}, \Delta) \propto \begin{cases} \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{1_j} \\ \frac{a+n_{j,-n}^{(i)}}{a+n_{j,-n}^{(i)}+b_{ij}^{(z)}} \prod_{t<i, t\in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } u_{1_j} < i < u_{L_j} \\ \prod_{t<i, t\in U_j} \frac{b_{tj,-n}^{(z)}}{a+n_{j,-n}^{(t)}+b_{tj,-n}^{(z)}} \cdot \frac{m_{j,-n}^{(d)}+\alpha}{m_{\cdot,-n}^{(d)}+K\alpha} & \text{if } i = u_{L_j} \\ 0 & \text{if } i \notin U_j \end{cases} \tag{3}$$

$$P(\Delta_{ij} = 1|\boldsymbol{\Delta}_{-i}, \mathbf{w}, \mathbf{z}) = \begin{cases} 0 & \text{if } n_j^{(i)} > 0 \\ \frac{\pi_{ij}}{\pi_{ij}+(1-\pi_{ij})\frac{B(a_{ij}^{(z)}, b_{ij}^{(z)})}{B(a,b)}} & \text{if } n_j^{(i)} = 0 \end{cases} \tag{4}$$

where z_{dn} is the subcommunity assignment for sequencing read w_{dn}^{i}. We define $n_{j,-n}^{(i)}$ as the number of times the *ith* taxa is assigned to the *jth* subcommunity and $m_{j,-n}^{(d)}$ as the number of times the *jth* subcommunity occurs in the *dth* sample, both excluding the current subcommunity assignment of z_{dn}. Additionally, we define $a_{ij}^{(z)} = a + n_j^{(i)}$ and $b_{ij}^{(z)} = b + n_j^{(i+1)} + ... + n_j^{(V-1)}$.

The chain is initialized with informative values for the z_{dn} variables by sampling from a multinomial distribution with taxa probabilities equal to the β_{ij} estimates from a standard LDA model. Once the chain has been run long enough to guarantee sufficient convergence, a set of the initial runs is removed as a "burn-in" period, and the remaining are taken as a set of samples from the target posterior distribution. As such, for each run, we can calculate estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as follows using the posterior predictive distribution:

$$\hat{\beta}_{ij} = P(w_{new}^{(i)}|z_{new}^{(i)} = j, \mathbf{w}, \mathbf{z}, \boldsymbol{\Delta}) = \begin{cases} \frac{a+n_j^{(i)}}{a+n_j^{(i)}+b_{ij}^{(z)}} & \text{if } i = u_{1_j} \\ \frac{a+n_j^{(i)}}{a+n_j^{(i)}+b_{ij}^{(z)}} \prod_{t<i, t\in U_j} \frac{b_{tj}^{(z)}}{a+n_j^{(t)}+b_{tj}^{(z)}} & \text{if } u_{1_j} < i < u_{L_j} \\ \prod_{t<i, t\in U_j} \frac{b_{ij}^{(z)}}{a+n_j^{(t)}+b_{tj}^{(z)}} & \text{if } i = u_{L_j} \\ 0 & \text{if } i \notin U_j \end{cases} \tag{5}$$

$$\hat{\theta}_j^{(d)} = P(z_{new} = j|\mathbf{z}) = \frac{m_j^{(d)} + \alpha}{m_{\cdot}^{(d)} + K\alpha} \tag{6}$$

The final estimate of $\boldsymbol{\theta}$ is defined as its posterior mean across all the runs. The final estimate of $\boldsymbol{\beta}$ can be found in a two-part process. First, calculate the posterior mean of Δ_{ij} across all runs, which is equivalent to a posterior estimate of π_{ij}. Then

dichotomize $\hat{\pi}_{ij}$ according to $I(\hat{\pi}_{ij} \geq 0.5)$. Next, assign $\hat{\beta}_{ij} = 0$ for any dichotomized $\hat{\pi}_{ij} = 1$, otherwise assign $\hat{\beta}_{ij}$ its respective posterior mean and normalize within each subcommunity such that $\sum_i \beta_{ij} = 1$.

3. RESULTS

3.1. Simulation Study

We conducted a simulation study to compare estimation accuracy and model fit between the proposed zinLDA and the standard LDA models. The data was simulated from a true zinLDA model, following the steps specified by the generative algorithm given section 2.3. First, we selected the total number of taxa (V) to be 120 across 150 independent microbial samples. Next, the total number of reads in each sample were drawn from a discrete uniform distribution with a lower bound of 5,000 and upper bound of 25,000. These parameters were selected to reflect real microbiome data sets aggregated to the genus-level classification. The number of subcommunities (K) was selected as five. The hyperparameter α of the Dirichlet distribution on $\boldsymbol{\theta}$ was set to $50/K$, as suggested for the original LDA model (Griffiths and Steyvers, 2004). Additionally, the hyperparameters π, a, and b of the zero-inflated generalized Dirichlet distribution on $\boldsymbol{\beta}$ were set to 0.4, 0.05, and 10, respectively. After running the simulation algorithm, the taxa that had a zero count for every sample, meaning a prevalence of 0%, were removed as such taxa would not be observed in a real data analysis. This reduced the total number of observed taxa (V_{obs}) to 87.

A zinLDA model with five subcommunities was fit to the simulated data set. Hyperparameters α, π, a, and b were set to their true values, as specified under simulation. Likewise, a standard LDA model with five subcommunities was fit, with default hyperparameter values of $50/K$ and 0.1 for α and η, respectively (Griffiths and Steyvers, 2004). To deal with the label switching problem commonly seen in Bayesian inference with mixture models, we use a method previously proposed to compare labels from an LDA model to their ground-truth. The pairwise Pearson correlation was calculated for each true-estimated subcommunity pair. The pair with the highest correlation is matched, then the pair with the next highest correlation among the remaining is matched, and so on until all true-estimated pairs are uniquely matched (Sankaran and Holmes, 2019).

To determine how well zinLDA is able to capture the latent community structure we compare the estimated β_{ij} for the top eight taxa per community to their true value and estimated value from the standard LDA model. **Figure 2** shows that both zinLDA and LDA correctly identify all of the top microbial taxa for each of the five subcommunities. Moreover, estimates from both models show low bias. We investigated how misspecification of the number of subcommunities influences zinLDA's ability to recover the representative taxa. An under-specified model, with one too few communities, collapses the representative taxa of two of the subcommunities together. Thus, resulting in both upwardly and downwardly biased estimates of β_{ij}, the taxa over subcommunity probabilities. The remaining three subcommunities have their representative taxa recovered and their respective β_{ij} estimates were not affected. Likewise, for an over-specified model, with one too many communities, it is able to accurately detect the five true subcommunity structures as specified under simulation, but identifies an additional nonsensical subcommunity that is composed of only one taxa (**Supplementary Figure 1**).

Fit of the two models was assessed through posterior predictive checks (Gelman et al., 1996). For each model, the posterior predictive distribution was used to simulate 100 data sets of the same dimensions as the original. The rationale behind using posterior predictive checks to assess model fit is as follows: if the model provides reasonable fit then the data simulated from the posterior predictive distribution, which is conditional on the observed data (X_{obs}) and the current model, should "look similar" to the observed data. We quantify how similar the observed data and the posterior predictive simulated data are by the test statistic $T(X) = X_{i\cdot}$, the count for the *ith* taxa. **Figure 3** plots the results from the posterior predictive checks. Each panel corresponds to a single biological sample. The y-axis plots $T(X)$ on the *asinh* scale. The x-axis plots each of the 87 taxa, ordered from smallest to largest based on the observed data for that sample. For large taxon counts we see that both models do well, with median values of both being similar to the truth, or observed, values. In contrast, we see that for small taxon counts the zinLDA model outperforms LDA. Specifically, for zero counts the zinLDA model is able to accurately estimate these counts better than its LDA counterpart. Across the 50 data sets simulated from the posterior predictive distribution, the zinLDA exhibits less over-smoothing for small taxon counts compared to the original LDA. Thus, this is an indication that the zinLDA model provides better fit to the data than the LDA.

To quantify how well the zinLDA model is able to distinguish between rare and absent taxa in each subcommunity we calculate the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We define a "positive" outcome as being a structural zero, $\Delta_{ij} = 1$, and a "negative" outcome as being a non-zero probability of belonging to that subcommunity, $\Delta_{ij} = 0$. The results show that under these simulation settings zinLDA can differentiate sampling and structural zeros with reasonable sensitivity and specificity (**Table 1**). Upon further examining the data, we noticed that the model we used to generate the data resulted in many taxa with very small true non-zero probabilities, making it very difficult to separate sampling zeros from structural zeros. To further demonstrate this point, we ran two additional simulations to see how different model parameters affect the posterior inference of being structural zeros. Both simulations reduce the number of taxa (V) to 50, but one also changes hyperparameter a, of the ZIGD distribution, to 0.5 from 0.05. **Table 1** shows that reducing the number of taxa without changing the value of a reduces the model's ability to differentiate between the two sources of zeros. In contrast, reducing V and also increasing a significantly increases the model's ability to accurately detect structural zeros, with such a modeling having sensitivity of 0.9 and PPV of 0.92. The sharp difference in the values of these diagnostic metrics between the models can be attributed to the fact that V, a, and b all influence the β_{ij} values, which in turn influences the probability of observing a sampling zero. For example, decreasing V without

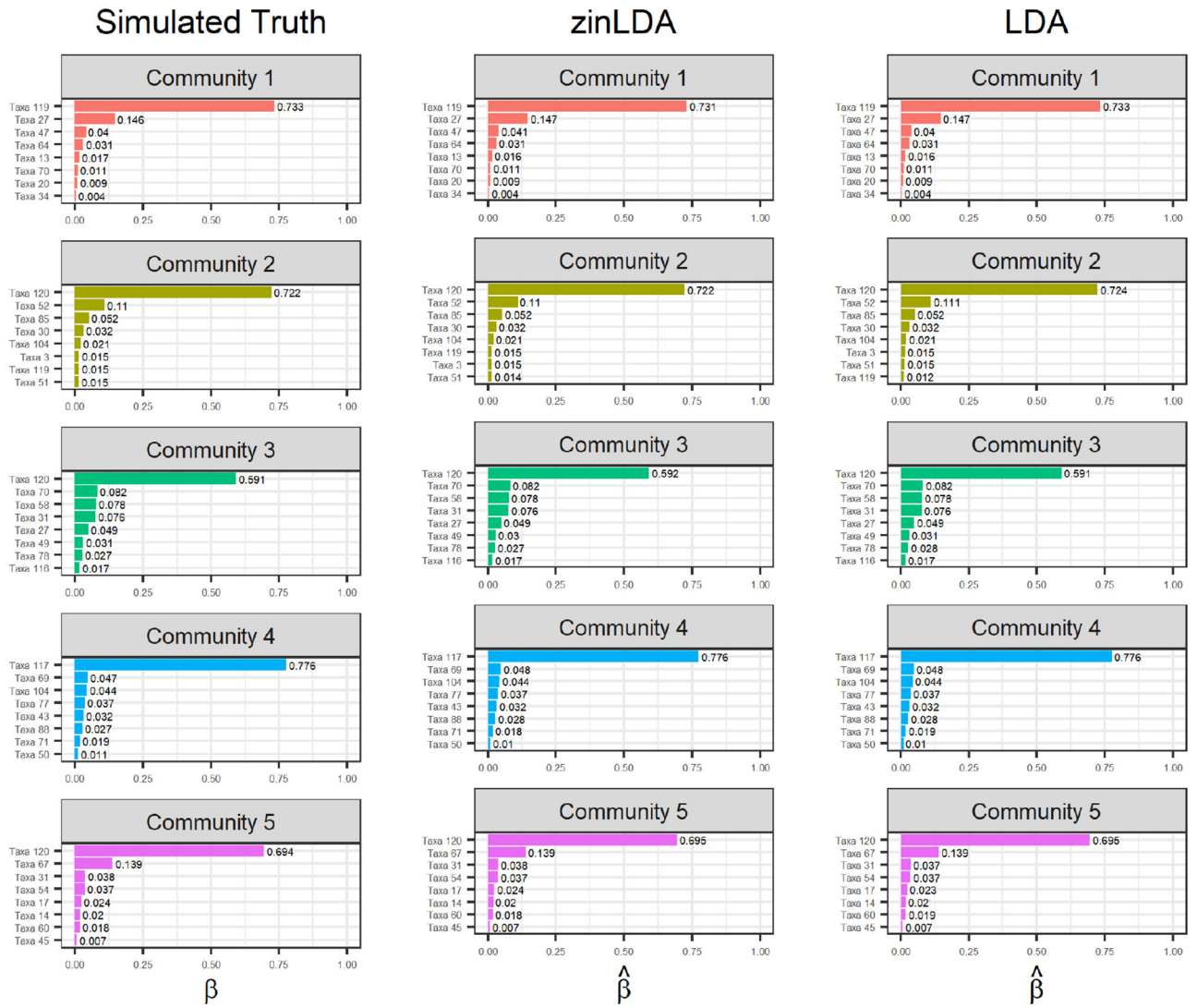


FIGURE 2 | Bar graphs of the top eight taxa, for each of the five sub-communities, with their corresponding β_{ij} values. The first column contains the "ground truth" top taxa from simulation. The second and third columns are the estimated top taxa from the zinLDA and LDA models, respectively.

changing a reduces many of the β_{ij} values, thus increasing the probability of observing a sampling zero. On the other hand, decreasing V in conjunction with increasing a increases many of the β_{ij} values and therefore decreases the probability of observing a sampling zero.

3.2. Real Data Applications

The American Gut Project (AGP) is a self-selected and open platform cohort. Citizen-scientists primarily in the United States, United Kingdom, and Australia, opted into the study, paid a fee to offset the cost of sample processing and sequencing, and gave informed consent (McDonald et al., 2018). All subjects provided a fecal microbiome sample and self-reported meta-data. The sequencing protocol used was identical to that of the Earth Microbiome Project (Gilbert et al., 2014; McDonald et al., 2018). The AGP microbial 16S rRNA gene sequencing data and meta-data are publicly available in The European Bioinformatics Institute repository under the accession ERP012803.

This analysis used a prior subset of the AGP data consisting of 3,679 subjects. Reads that were ambiguously assigned or unassigned at the genera level were removed. Moreover, genera with a prevalence of <20% across all samples were removed. After this filtering of the microbial genera, any samples with a total number of reads of zero were removed. This left 3,566 samples and 70 unique genera for downstream analyses.

A random subset of 1,000 subjects from the AGP data was sampled, a zinLDA model with five subcommunities and hyperparameter values being specified the same way as in the simulation study was fit. When possible, the choice of the number of latent subcommunities should be informed by biological or clinical reasoning. In the absence of such, data-driven approaches may be used. In particular for the

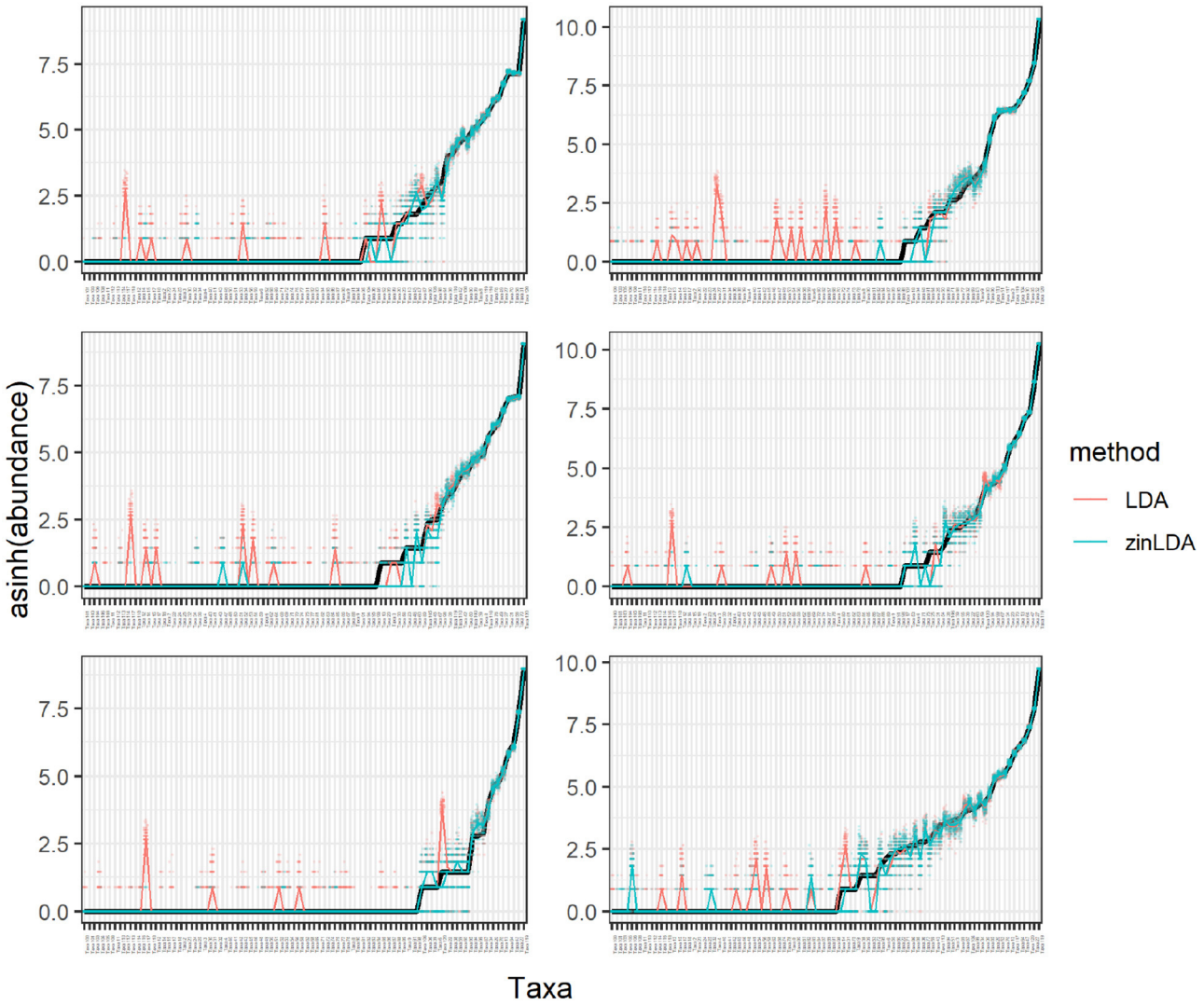


FIGURE 3 | Observed and posterior predictive simulated *asinh*-transformed taxon counts plotted in order of increasing true abundance. Each panel is a different biological sample. The solid black line represents the true counts. The pink and blue points are the counts from 50 posterior predictive simulated data sets and the pink and blue solid line represents the median counts across all 50 data sets, from the LDA and zinLDA models, respectively.

AGP data, K was determined by comparing the log-likelihood, AIC, and representative taxa across many models, each with a different number of subcommunities, applied to a set of 1,000 independently selected subjects. These results were robust across slight changes in the number of subcommunities.

The representative taxa from each subcommunity and their membership probability (β_{ij}) is shown in **Figure 4**. We observe that each subcommunity is characterized by one single dominant taxa, including *Faecalibacteruim*, *Prevotella*, *Bacteroides*, *Acinetobacter*, and *Akkermansia*.

Model fit was assessed via posterior predictive checks and compared to that of the standard LDA model. Since current sequencing technology, such as 16S rRNA gene sequencing, can only provide quantification about relative abundance model fit was assessed using both the relative abundance and the observed counts (**Supplementary Figures 2, 3**). The two plots exhibit similar patterns, indicating the difficulty in fitting the small count data. Another explanation of observing such similar model fits is that our analysis did not identify structural zeros

TABLE 1 | Comparison of estimated structural zero taxa from the zinLDA model to true structural zero taxa from simulation across different parameter settings using sensitivity, specificity, positive predictive value, and negative predictive value.

	Sensitivity	Specificity	PPV	NPV
$V = 50, a = 0.5$	0.90	0.94	0.92	0.93
$V = 50, a = 0.05$	0.51	0.51	0.40	0.61
$V = 87, a = 0.05$	0.73	0.67	0.59	0.79

A "positive" results is assumed to be $\beta_{ij} = 0$.

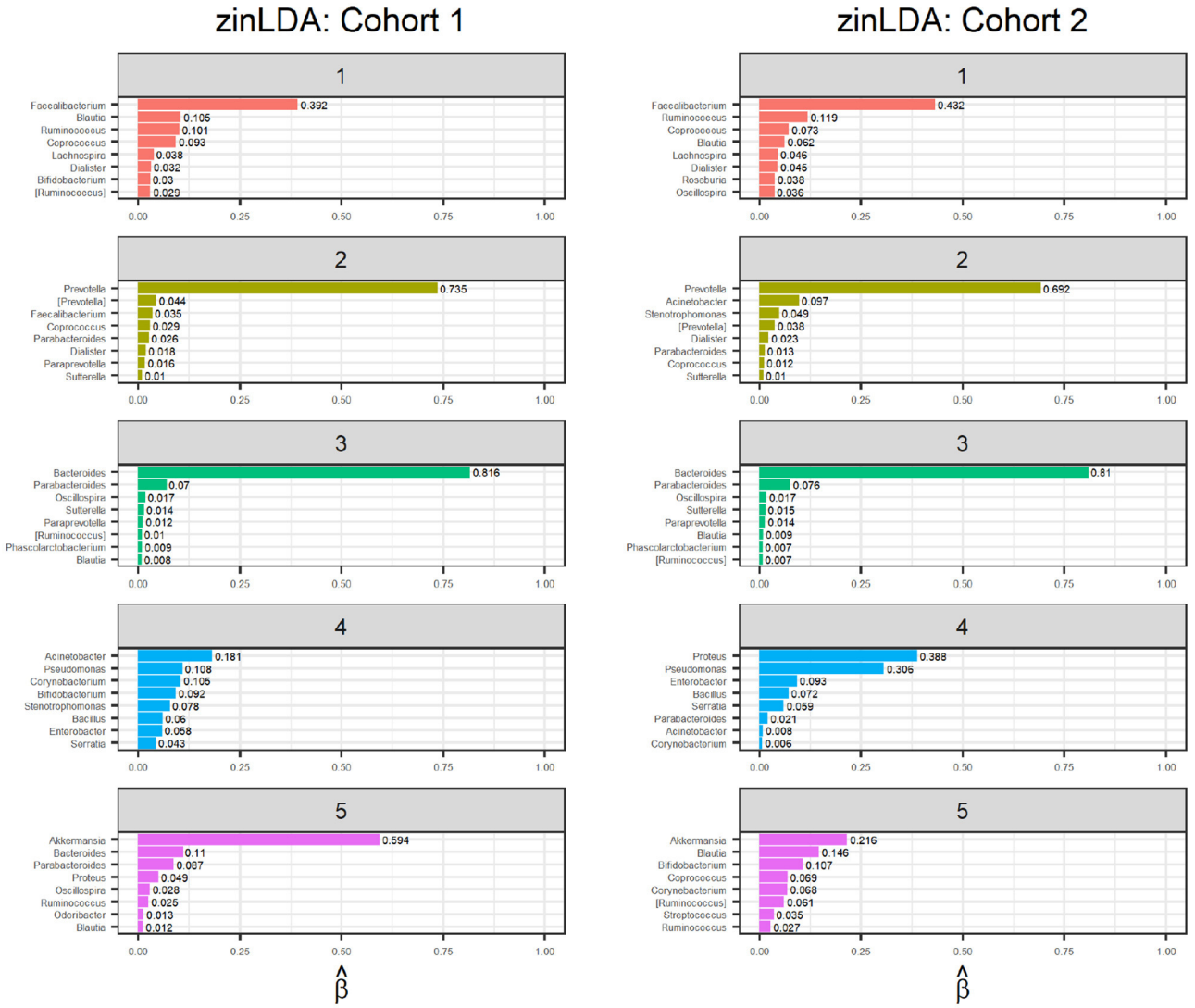


FIGURE 4 | Bar graphs of the representative taxa for each of the five sub-communities, with their corresponding β_{ij} values, for two independent subsets of 1,000 randomly selected subjects from the American Gut Project.

with strong evidence in our data. **Supplementary Figure 4** shows the posterior estimates of the probability of zero count being a structural zero for each of the taxa in each subcommunity, indicating relatively weak evidence of being structural zeros.

Finally, to determine whether the model is stable, meaning it detects true subcommunity clusters of co-occurring taxa and is not clustering the noise in the observations, we apply an identical zinLDA model to another set of 1,000 AGP microbial samples that is independent from the first. The representative taxa from this validation set is compared to that of the first cohort (**Figure 4**). The subcommunities between the two cohorts were matched using pairwise correlations as done in simulations. The average cosine similarity of the matched subcommunities is 0.80. The results show that the communities identified by zinLAD are very stable and replicable.

4. DISCUSSION

The micro-organisms that constitute the human microbiome form subcommunity-like structures via dynamic and complex interactions with one another. Identifying these structures is imperative for a better understanding of how these microbes influence human-host health. We propose a zero-inflated latent Dirichlet allocation model, a further modification of the LDA model that amounts to changing the prior distribution on the taxa per subcommunity distribution to a zero-inflated generalized Dirichlet from a Dirichlet distribution. Despite this change our model retains the advantageous conjugate prior property between the ZIGD and multinomial distributions. As such, we are able to implement an efficient Gibbs sampling algorithm, with only one additional step compared to that of LDA, for parameter estimation.

zinLDA modifies the LDA model proposed by Blei et al. (2003) to allow for subcommunities to be composed of a subset of all the microbes in a cohort of samples. Mathematically, since a subcommunity is defined as a distribution over taxa, this is equivalent to assigning some taxa a zero-probability of belonging to it. This is particularly advantageous in microbial analyses as it allows for a clear distinction between sampling and structural zeros within a subcommunity structure. Structural zeros come from those zero-probability taxa; they are truly absent from the community. Sampling zeros come from taxa that do belong to the community, but with low probability, and thus were not captured due to shallow sequencing depth. Due to this adjustment, zinLDA model can be used to simulate more realistic sparse count data than models such as the Dirichlet multinomial or Dirichlet multinomial mixture models.

We used simulation studies to compare the two models and investigate where zinLDA outperforms the standard LDA model. First, we show that the two performed equally well in identifying the representative taxa for each subcommunity. This is to be expected as the LDA model already does a good job in identifying common taxa and the zinLDA estimates of the community assignment for each sequencing read were initialized using the results from a standard LDA model. The performance gain in using zinLDA is seen when examining the low probability and absent taxa within in each subcommunity. The greatest performance gains are made when the probability of being a sampling zero is not too small. Furthermore, we use real data from the citizen scientists of the American Gut Project to show that our method can detect potentially meaningful biological and ecological subcommunities of microbial species. By assigning each sample a probability of belonging to each of these subcommunities we are also able to gather information about population level microbial structures.

As for any Bayesian models, zinLDA requires the hyperparameters to be pre-specified. In our analysis of the real data sets, we used the same hyperparameters as in our simulations and explored various other choices. For the same number of communities, we observed that the community structures and the representative taxa were not too sensitive to the values of these hyperparameters. Determining the number of clusters or subcommunities is a hard problem, as for any clustering methods. For real data analysis, we suggest that the users try different numbers of K, evaluate the sub-community structures, and then choose one based on both the sizes of the communities and also possible biological interpretations.

Finally, the zinLDA model can be used to simulate more realistic microbiome count data that allow for both structural zeros and sampling zeros. Such simulations can be used to evaluate various statistical tests developed for microbiome data analysis, including evaluating power of the tests for differential abundance and methods for modeling microbiome count data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets for this study can be found in EBI under project PRJEB11419 and Qiita study ID 10317. Software for implementing the method described in this manuscript is publicly available on GitHub at https://github.com/rebeccadeek/zinLDA.

AUTHOR CONTRIBUTIONS

RD and HL developed the ideas and the methods together, analyzed the real data sets, and wrote the manuscript. RD implemented the methods and performed the numerical analysis.

FUNDING

This research was funded by NIH grants GM123056 and GM129781. NIH funded these projects to develop computational methods for analysis of microbiome and metagenomic data. The grants will also cover the open access publication fees.

ACKNOWLEDGMENTS

We thank the participants of the American Gut Project for sharing their data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.602594/full#supplementary-material

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.5555/944919.944937

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14, 244–258. doi: 10.1093/biostatistics/kxs038

Connor, R., and Mosimann, J. (1969). Concepts of independence for proportions with a generalization of the dirichlet distribution. *J. Am. Stat. Assoc.* 64, 194–206. doi: 10.1080/01621459.1969.10500963

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* 6, 733–807.

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The Earth microbiome project: successes and aspirations. *BMC Biol.* 12:69. doi: 10.1186/s12915-014-0069-1

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl. 1), 5228–5235. doi: 10.1073/pnas.0307752101

Hawinkel, S., Kerckhof, F.-M., Bijnens, L., and Thas, O. (2019). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLoS ONE* 14:e0205474. doi: 10.1371/journal.pone.0205474

Higashi, K., Suzuki, S., Kurosawa, S., Mori, H., and Kurokawa, K. (2018). Latent environment allocation of microbial community data. *PLoS Comput. Biol.* 14:e1006143. doi: 10.1371/journal.pcbi.1006143

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* 7:e30126. doi: 10.1371/journal.pone.0030126

Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M., and Hamada, M. (2020). Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. *Microbiome* 8:95. doi: 10.1186/s40168-020-00864-3

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.* 25, 217–228. doi: 10.1016/j.tim.2016.11.008

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3. doi: 10.1128/mSystems.00031-18

Raman, A. S., Gehrig, J. L., Venkatesh, S., Chang, H.-W., Hibberd, M. C., Subramanian, S., et al. (2019). A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science* 365:6449. doi: 10.1126/science.aau4735

Sankaran, K., and Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostatistics* 20, 599–614. doi: 10.1093/biostatistics/kxy018

Sommeria-Klein, G., Zinger, L., Coissac, E., Iribar, A., Schimann, H., Taberlet, P., et al. (2020). Latent Dirichlet allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Mol. Ecol. Resour.* 20, 371–386. doi: 10.1111/1755-0998.13109

Tang, Z.-Z., and Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* 20, 698–713. doi: 10.1093/biostatistics/kxy025

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/fgene.2021.612670

Non-linear Normalization for Non-UMI Single Cell RNA-Seq

Zhijin Wu[1]*, *Kenong Su*[2] and *Hao Wu*[3]

[1] Department of Biostatistics, Brown University, Providence, RI, United States, [2] Department of Computer Science, Emory University, Atlanta, GA, United States, [3] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, United States

Single cell RNA-seq data, like data from other sequencing technology, contain systematic technical noise. Such noise results from a combined effect of unequal efficiencies in the capturing and counting of mRNA molecules, such as extraction/amplification efficiency and sequencing depth. We show that such technical effects are not only cell-specific, but also affect genes differently, thus a simple cell-wise size factor adjustment may not be sufficient. We present a non-linear normalization approach that provides a cell- and gene-specific normalization factor for each gene in each cell. We show that the proposed normalization method (implemented in "SC2P" package) reduces more technical variation than competing methods, without reducing biological variation. When technical effects such as sequencing depths are not balanced between cell populations, SC2P normalization also removes the bias due to uneven technical noise. This method is applicable to scRNA-seq experiments that do not use unique molecular identifier (UMI) thus retain amplification biases.

Keywords: scRNA sequencing, single cell, normalization, statistical method, gene expression

OPEN ACCESS

Edited by:
Mengjie Chen,
University of Chicago, United States

Reviewed by:
Himel Mallick,
Merck, United States
Xiang Zhou,
University of Michigan, United States

***Correspondence:**
Zhijin Wu
zhijin_wu@brown.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *30 September 2020*
Accepted: *05 March 2021*
Published: *09 April 2021*

Citation:
Wu Z, Su K and Wu H (2021) Non-linear Normalization for Non-UMI Single Cell RNA-Seq. Front. Genet. 12:612670. doi: 10.3389/fgene.2021.612670

1. INTRODUCTION

Single Cell RNA-sequencing (scRNA-seq) has become a widely applied tool to study the diverse and dynamic transcriptional activities among cell populations (Tang et al., 2009). Before the RNA-sequencing technology was applied to query the transcriptomes of individual cells, scientists have used it widely to measure mRNA expression from bulk samples (Mortazavi et al., 2008), in which an average level of RNA expression from a large number (often millions) of cells is obtained. Methods for data processing, including mapping short reads to the reference transcriptome and normalization to account for technical variability in the efficiency of RNA extraction, amplification and counting, evolved along the progress of the sequencing technology. These include simple size factors to adjust for global effects such as sequencing depth, such as widely used count per million (CPM) or reads per million per kilobase (RPKM) for their simplicity (Mortazavi et al., 2008), and more data adaptive trimmed mean of M values (TMM) (Robinson and Oshlack, 2010). Noting that non-linear and inconsistent biases due to gene length and GC-content exist in RNA-seq data, more flexible methods have been proposed, such as the conditional quantile normalization (CQN) (Hansen et al., 2012) and remove unwanted variation (RUV) (Risso et al., 2014).

All normalization methods, explicitly or implicitly, make assumption about characteristics of the data that are expected. For example, in many bulk RNA-seq data sets, assumptions on the lack of global shifts of the distribution of expression are often reasonable. As a result, the changes of the location, scale, or shape of the distribution are attributed to technical effects and removed in normalization (Robinson and Oshlack, 2010; Hansen et al., 2012). scRNA-seq data share many

similarities of bulk RNA-seq data, but have their unique characteristics. These include, but are not limited to, the much higher percentage of genes with zero count and generally lower library size (Shapiro et al., 2013). In addition, there is often much greater variability among cells compared to that among bulk samples, because bulk samples measure the average expression from a large population of cells (Wu et al., 2014). Thus, it may no longer be reasonable to assume the lack of global differences, and a direct adaptation of bulk RNA-seq normalization is not optimal, despite its convenience.

The need for specialized normalization is well-recognized. Since the introduction of scRNA-seq, a handful of normalization approaches have been proposed (Lun et al., 2016; Bacher et al., 2017). Most analyses of RNA-seq data at least attempt to address this bias due to sequencing depth or overall mRNA capture efficiency by turning the counts data into counts-per-million (CPM). This practice implicitly assumes a linear relationship between library size and the observed counts. There are several problems with this simple practice. One is that the library size (the total observed count in a sample) may not be a stable statistic to represent the overall counting efficiency in a cell. In bulk RNA-seq, each individual gene accounts for a very small fraction of a sample, thus the library size often captures the overall efficiency including sequencing depth and mRNA extraction efficiency. In scRNA-seq, a few top genes can account for a large fraction of total counts, making the library size sensitive to the variation of these genes, which are not necessarily stable across cells. This problem can be alleviated when one uses a more robust estimate of the size factor, such as using TMM. Another issue with a simple size factor adjustment is that it assumes the impact of the size factor is the same to all genes in the same cell. Bacher et al. (2017) showed that this is not necessarily true, and proposed to normalize genes in several groups. Recognizing that common assumptions on an identical distribution of genes expression may not be reasonable across all cells, normalization based on internal ERCC controls have also been proposed (Ding et al., 2015). However, since the control RNAs are spiked in after RNA extraction, the ERCC controls only capture technical biases in a portion of the sample preparation procedures. Though 96 RNAs are included in the ERCC panel, many of them are at levels too low to be detected, making the number of controls that can be used to capture the systematic bias much lower, thus the biases less reliably estimated.

In this manuscript, we describe a simple but effective normalization procedure that captures the potential non-linear, systematic biases in scRNA-seq data. We consider that a gene's observed count is affected by both its expression level (the biological factor) and the detection efficiency (the technical factors). The technical factors include the quality of cell dissociation, mRNA extraction/amplification efficiency, and sequencing depth. These factors may have different impact across genes. The combined effect of these factors on detection efficiency is the technical bias we aim to estimate and remove. Our procedure takes into account both gene-specific and cell-specific contexts in scRNA-seq data, thus borrows information both from the same gene across cells and from other genes within the same cell to achieve a robust normalization factor.

2. RESULTS

2.1. Data Sets

We use four scRNA-seq data sets to illustrate the normalization performance. The first is from a type 2 diabetes study of pancreatic islet cells, referred to as "T2D" data hereafter. The T2D data set includes 978 cells, of which 239 are alpha cells (Lawlor et al., 2017). We use the alpha cells as an example to illustrate variation within a cell type. This data set is available at Gene Expression Omnibus (GEO) with accession number GSE86473. The second data set (GEO accession number GSE85917) profiles human embryonic stem cells, referred to as "hESC" data hereafter. There are 92 H1 cells sequenced twice with very different sequencing depth: approximately one and four million reads per cell. This data set was originally generated to evaluate SCnorm normalization method (Bacher et al., 2017). The third data set (GEO accession number GSE45719) profiles cells in different early development stages ranging from zygote to blastocyst and is referred to as the "embryo" data using Smart-seq (Deng et al., 2014). The fourth data set (GEO accession number GSE75748) comes from a time course experiment that measured hESC cells at different time points, including 758 cells, and is referred to as the "time course" data (Chu et al., 2016).

2.2. The Technical Bias May Not Be a Constant Linear Effect of Library Size

The impact of overall mRNA extraction efficiency and sequencing depth is well-known. In single cell data this is reflected in two ways: cells with higher library size tend to have higher gene detection rate (the proportion of genes with non-zero count), and tend to have higher counts on the genes that are observed. The simplest adjustment for this overall effect is turning the counts data into counts-per-million (CPM). This practice inexplicitly assumes a linear relationship between library size and the observed counts, and makes the same adjustment for all genes in a given cell. We first demonstrate that technical bias depends on the gene as well, and is not always a simple linear effect.

For cell i, denote the library size by L_i. Consider gene g in this cell, denote its gene expression level as θ_{gi}, and the observed read count as Y_{gi}. When we assume that $E[Y_{gi}] \propto \theta_{gi}L_i$, normalizing by Y_{gi}/L_i is a reasonable practice. This type of normalization, using a cell-wise size factor, implies $\log(E[Y_{gi}]) = \log(\theta_{gi}) + \log(L_i) + c$. It means that the log transformed counts are proportional to log library size with a constant slope 1 for all genes. We explore these assumptions in real scRNA-seq data as shown in **Figure 1**, where we plot the slope of log counts regressing on library size against the correlation between a gene's counts and library size across cells. If we had a constantly expressed gene with $\theta_{gi} \equiv \theta_g$ and the gene counts are proportional to L_i, we would have a perfect correlation and slope 1. Here we focus on genes that are reliably detected and only include those with average log counts greater than 4. As expected, the counts for many genes are strongly correlated with library size, confirming that the library size indeed affects measured expression level, though the correlation is lower than 1 since there are natural variations of expression levels even within the same cell type. The correlation

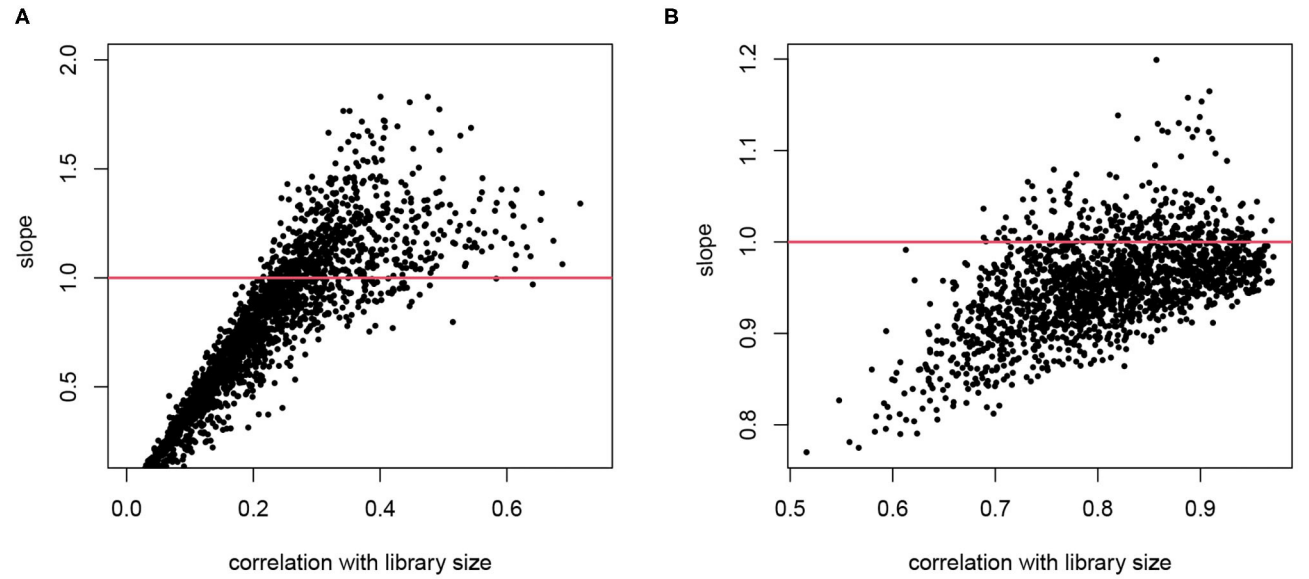


FIGURE 1 | The relationship between counting efficiency with library size. The log transformed counts are regressed on log library size and the slope is plotted against the correlation between the gene counts and the library size. A slope of 1 indicates a linear effect of library size on counting efficiency. **(A)** Data from T2D study. **(B)** Data from hESC study. Only genes with average log counts greater than 4 are included.

with L_i is lower for genes with high biological variation or genes with low expression and hence under greater influence of Poisson counting error. The slopes from genes that are highly correlated with library size are the most informative of the extent of the technical bias. We observe that the assumption of a constant slope of 1 is inaccurate in two senses: (1) the slopes between $\log(Y_{gi})$ and $\log(L_i)$ are not necessarily the same for all genes; and (2) the slope on average is not necessarily 1. In the T2D data, the slope tends to exceed 1 for genes that show high correlation with library size, whereas in the hESC data the slope tends to be lower.

2.3. Not All Genes Reflect Technical Bias in a Cell

Bacher et al. (2017) report similar observations that the need for normalization differs for different genes and give specific examples of genes with high, median and even negative slope in this relationship in the data used in **Figure 1B**. As a solution, they divide the genes into multiple bins and estimate their "count-depth relationship" separately, and normalize accordingly.

We take a different approach here without putting genes into bins. Instead, we obtain a cell- and gene-specific normalization factor that depends on the mean expression level, represented by a smooth function. This is motivated by the fact that most, if not all, genes are not transcribed in all cells. When a gene is expressed, we often observe a close-to-linear relationship between the gene count and the library size, as seen in **Figure 2**. This means that a higher count observed could be a result of higher sequencing depth or higher mRNA extraction success in certain cells, instead of higher expression level. This is the motivation behind CPM type of normalization. However, we also notice that even in cells with very high library size, we often observe low but non-zero counts, shown in red in **Figure 2**. We have introduced a two-phase expression model, *SC2P*, for scRNA-seq data that account for these two latent phases (Wu et al., 2018). Phase I corresponds to a background level of counts which represent the inactive phase, and Phase II corresponds to the phase when the gene is actively transcribed. For a cell that has high extraction/amplification rate and is sequenced deeply, the active genes in it tend to show higher counts. In the same cell, genes in Phase I will only have a low, background level of counts, regardless of the library size.

2.4. Technical Bias Depends on Expression Level

The variation in gene counts is a combined result of biological variation, which we desire to retain, *systematic* technical variation, which we aim to remove in normalization, and lastly, random noise, which is not identifiable from the biological variation. In **Figure 3**, we illustrate an example of the systematic bias manifested differently in the two latent phases. This figure is similar to the "MA plot" commonly used in gene expression microarray data. Here, each point represents a gene. The x-axis is the mean expression within a given cell type, and the y-axis is the log ratio of a gene's count in this particular cell versus the mean expression level. This plot shows the overall pattern of bias as a function of expression level. A symmetrical scatter of points around the $y = 0$ line reflects no need for normalization. A simple linear effect of the library size leads to a constant bias in the log scale, hence the points shift vertically, and will be symmetrical around $y = \log L_i - \log L_0$ for sample i, where L_i and L_0 are the library sizes for the specific cell and the reference (typically set to be the median library size in a data set). However,

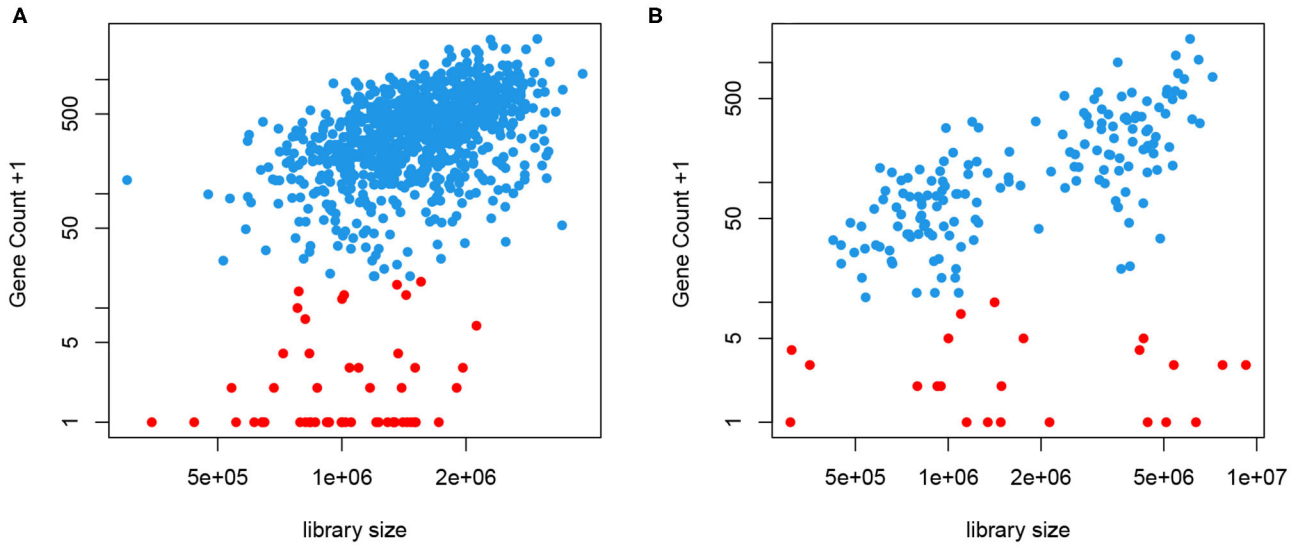


FIGURE 2 | Gene counts are positively correlated with library size when the gene is expressed (blue) but appear to have little correlation with library size when they are in the background phase (red). **(A)** Data from T2D study. **(B)** Data from hESC study.

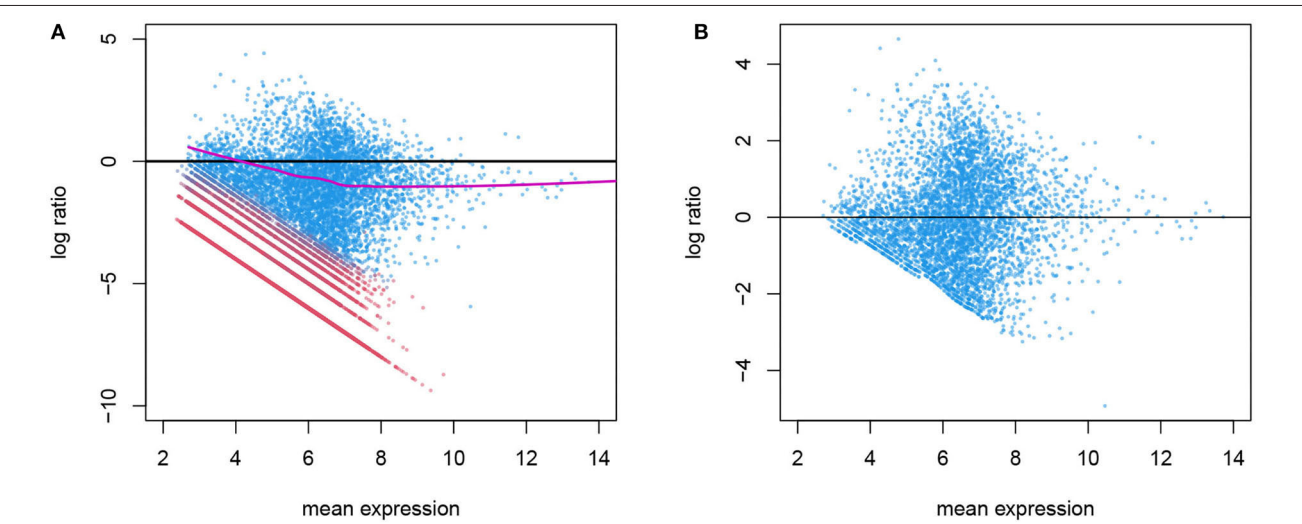


FIGURE 3 | (A) Before normalization MA plot for gene counts before normalization. The log ratio of a gene's count in one cell over its mean expression level is plotted against the log mean expression (the alpha cells in the T2D data are used here). Genes in active expression phase are shown in blue and genes in background phase shown in red. Only the blue genes are used in estimating the systematic bias, shown in purple. **(B)** After normalization: the log ratio from normalized genes in the active phase shows no systematic bias.

sometimes the bias depends on the expression level and cannot be captured by one constant, and a non-linear normalization is needed. This has been used for diagnosis as well as for estimating and removing the systematic bias in microarray data (Bolstad et al., 2003). One key difference is that in scRNA-seq data, not all genes in a cell are affected by the systematic bias to the same extent. As shown in **Figure 2**, a gene's count is affected only when it is in the active phase. Thus, counts from genes who are in the background phase do not contain information about the sequencing efficiency, and should not be included in the estimation of the systematic bias.

In Wu et al. (2018) we show that the distribution of background counts and that of genes in the active phase are cell- and gene-specific, so a universal cutoff to determine the phase is not ideal. We describe a mixture model using a zero-inflated Poisson distribution and a lognormal-Poisson distribution for the two phases and estimate the conditional probability that a gene is in the active phase, given its gene identity and the cell context. This allows us to divide the counts in a cell to the two phases as shown in **Figure 3**. The systematic bias due to inconsistent sequencing efficiency can then be estimated as a smooth curve using the gene counts in the active phase alone.

2.5. Removing the Count-Depth Dependence

The goal of normalization procedures is to remove technical variability without removing biological variability. One indication of unwanted technical variability is that gene counts are positively correlated with library size, referred to as the count-depth relationship (**Supplementary Figure 2A**). After adjusting for size factors, this strong correlation is often reduced toward zero, as seen in **Figure 4** and **Supplementary Figures 2B–D**, since many normalization factors directly aim to remove the library size effect. However, we also notice that negative correlation is often introduced to genes with lower average expression levels in simple global normalization approaches, indicating an over-adjustment for those genes. SCnorm and SC2P both reach a near 0 correlation overall, with the result from SC2P closer to zero for genes over a wider range of mean expression level. **Supplementary Figure 3** reveals the similarity and difference between SC2P and SCnorm more directly by plotting the raw and normalized counts in the same cell. We see that both methods adjust the higher counts even higher, but lower counts to a lesser extent. SCnorm partitions genes into several groups, each forming a curve, with different levels of adjustment. SC2P does the adjustment in a smooth fashion without putting genes in discrete categories, thus lacking apparent clusters in the figure.

2.6. Removing Technical Variation and Maintaining Biological Difference

To show the success in removing technical variations, we first compare the conditional standard deviation of gene expression levels. Since dropout is a common phenomenon in scRNA-seq data, even strong cell type marker genes are not always observed in the corresponding cell type. Thus, marginal standard deviations often obscure the actual variability (**Supplementary Figure 1**). For each gene, we compute the standard deviation of its expression level when the gene is reliably detected, based on the posterior probability of a gene in the active phase. Among cells of the same type, we expect that the variance has sources of both biological and technical origins, and we expect that the variance reduces in normalized data. To evaluate the reduction in variance we compute the ratio of the variance in the normalized versus raw data. In **Figure 5A** we compare the ratio in genes stratified by average expression levels, in Alpha cells from the T2D data. Several methods (SCnorm, scran, and SC2P) can reduce the variance in highly expressed genes. Many, however, lead to an increase of variation for genes with lower expression levels. SC2P is the only method that can reduce the variance throughout the entire range of mean expression. In this particular data set, the normalization in DESeq actually increased the variance.

We certainly want to make sure that we do not reduce signal in the process of removing technical variation. To confirm this we show the difference in average expression between the Alpha and Beta cells. As shown in **Figure 5B**, the log fold change computed in SC2P normalized data maintains the between cell type differences. Similar results from the embryo data are included in the **Supplementary Figure 4**.

2.7. Removing Bias Due to Unbalanced Technical Bias

When the technical biases are randomly and evenly distributed in two cell populations, the population mean expression suffers from much smaller bias than the expression level in individual cells, since the law of large numbers will make the average

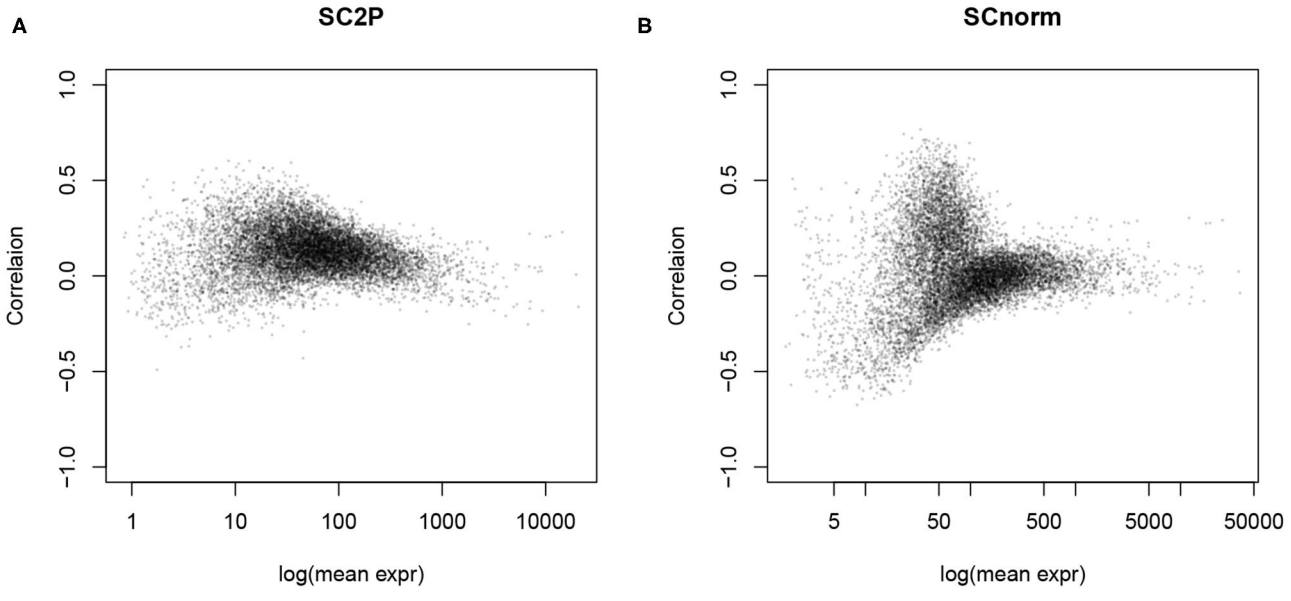


FIGURE 4 | The count-depth relationship, measured as the correlation between the (normalized) gene counts and library size, is reduced after normalization. **(A)** Normalized by SC2P normalization. **(B)** Normalized by SCnorm.

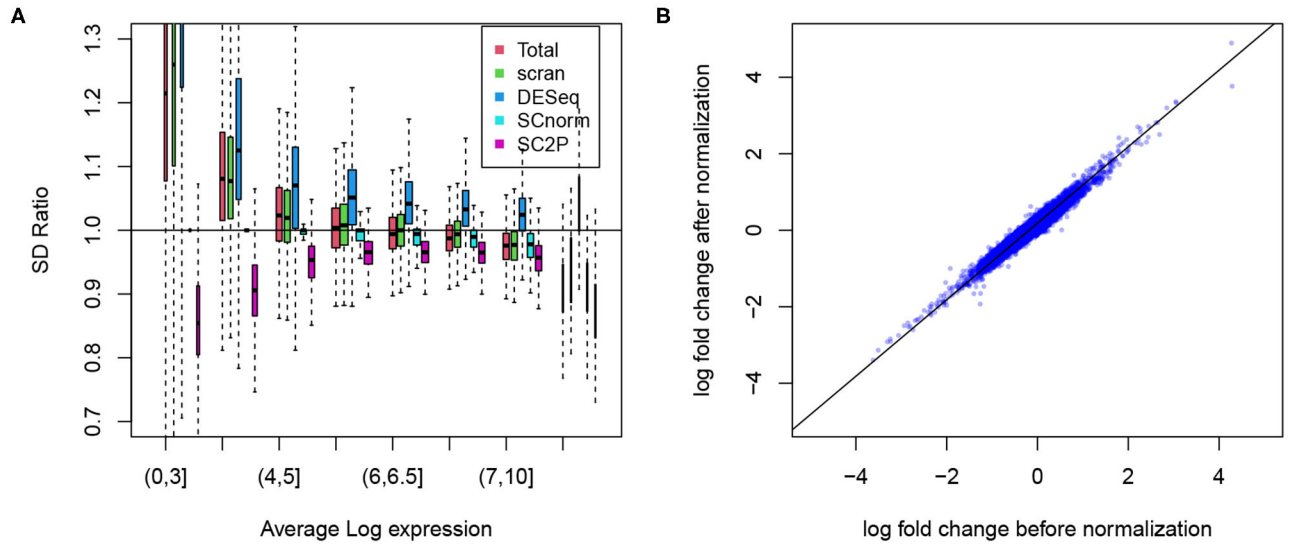


FIGURE 5 | (A) Reduction of the technical variability among replicate cells. The ratio of gene specific standard deviation in normalized log counts over that in the raw log counts plotted. Genes are displayed in different groups based on their average expression when they are expressed. **(B)** The log fold change between Alpha and Beta cell populations before and after normalization remains at similar levels.

of technical noise converge to zero when the number of cells increases. However, when two populations of cells in comparison have different distributions of technical effects, we may have biased result even in population means. For example, if one cell population tends to have more deeply sequenced cells than the other cell population, we will observe a bias in the mean expression levels, and DE observed across the two groups may simply reflect the imbalance in sequencing depth in the two populations. Successful normalization should remove such biases without introducing new biases.

For illustration purpose, we use the hESC data set that profiles H1 cells with both high and low sequencing depth so the systemic bias is obvious. When the sequencing depth is unbalanced between the two groups, the group with more highly sequenced cells tend to have average expression biased up, creating positive log fold change in genes without true DE. Here we compare the ability of various normalization methods in their ability to remove this potential bias. **Figure 6** shows the boxplots of log fold changes of normalized gene expression for a two-group (the same type of cells in high- vs. low-sequencing depth groups) comparison, where the genes are stratified by average expressions. Since there is no biological difference between the two groups, we expect the log fold changes to be around zero. We see that, for highly expressed genes, all methods appear to remove the technical bias and show a median at zero. For lower expressed genes, the normalization methods using a cell-wize normalization factor (Total, scran, and DESeq) actually introduce biases to the data. This is because the lower expressed genes are affected by the library size in a lesser degree, thus they are over-normalized. SCnorm, by normalizing genes in different groups, can alleviate this problem to some extent and show smaller bias after normalization. SC2P is the only normalization that works well for genes with different average expression levels.

2.8. Impact on Downstream Analysis

The flexibility and single cell resolution of the scRNA-seq technology lead to a wide variety of applications and a large number of new analysis methods. To illustrate the consequences of normalization procedures on downstream analysis, we present two examples below. The first is differential expression analysis. Due to the lack of biological ground truth, we do not directly compare the accuracy of DE magnitude or the sensitivity of DE detection. Instead, we assess the impact of normalization on the robustness of DE detection. In scRNA-seq, the number of cells in each population is often orders of magnitude higher than the number of samples in most bulk RNA-seq data. A robust and reproducible analysis should not have results that are sensitive to the inclusion or removal of a few cells. We illustrate with the time course data and compare expression between time points We show that different normalization methods lead to different reproducibility in the time course data. When 5 cells, either the ones with the highest library size, or randomly chosen, are removed from the data set, our normalization shows much less disruption. In contrast, data normalized with other alternatives could lead to drastic changes (**Supplementary Figure 5**).

We also compare the impact on clustering using the embryo data. We use log transformed pseudo counts after different normalization in three widely used scRNA-seq clustering methods, including SIMLR ((Wang et al., 2017), SHARP (Wan et al., 2020), and SC3 Kiselev et al. (2017). **Figure 7** compares the Adjusted Rand Index (Hubert and Arabie, 1985), which measures the concordance of pair-wise relationship between each pair of cells with known developmental stages, adjusted for the agreement due to coincidence. The proposed normalization has the highest ARI in all three methods.

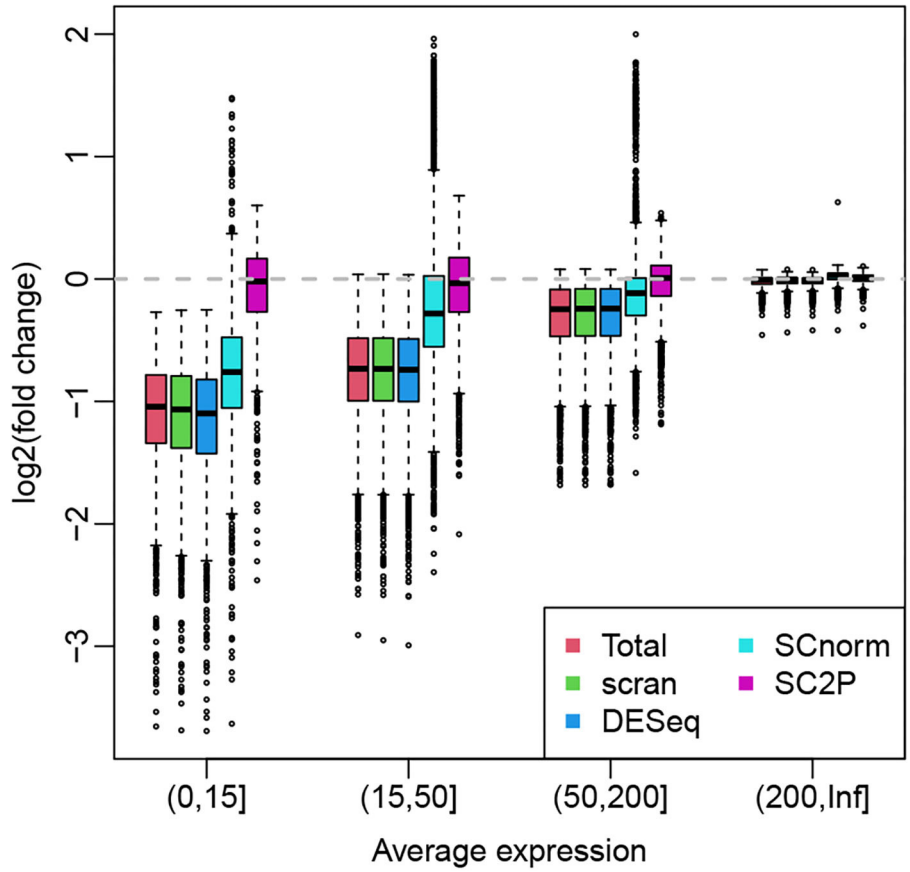


FIGURE 6 | Box plots of log fold changes of normalized gene expression from different methods. The genes are stratified by average expressions. The log fold changes are computed based a two-group comparison, where the two groups contains cells with high and low sequencing depths. All cells are H1 hES cells.

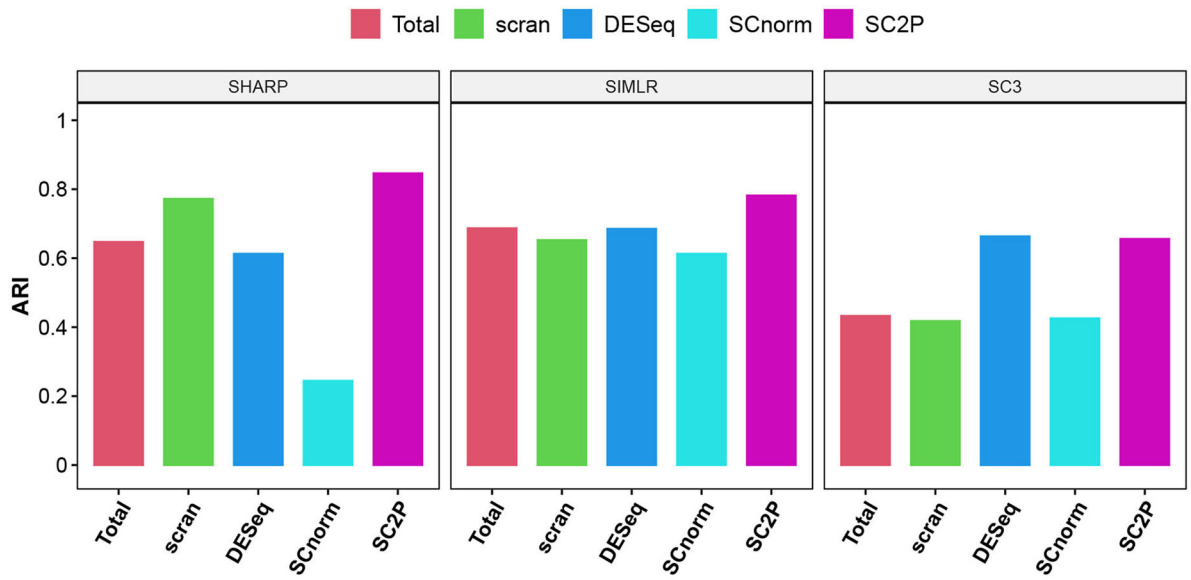


FIGURE 7 | The adjusted Rand index (ARI) using three different clustering methods with different normalization. Genes that are detected in at least three cells are used in clustering.

3. DISCUSSION

We present a normalization method that provides a cell- and gene-specific normalization factor that borrows information across genes and across cells. Both the cell context and gene context are used in predicting whether a gene appears to be in the active phase in a given cell, and only the active ones are used in estimating the technical bias due to RNA extraction/amplification/sequencing. It is more flexible than simple size factor normalization, which adjusts all genes in a cell in a universal manner, but is still robust for the normalization is estimated from a large number of genes using only a few degrees of freedom.

scRNA-seq opens the door to many new applications beyond what is offered by bulk RNA-seq. It allows the query of the heterogeneity of individual cells, instead of the average of many. This means higher variability of the direct measurements, since the quantity measured is no longer a population average which is stabilized when millions of cells are pooled together. This often means that we have many more cells sequenced in an experiment, thus many more "samples" to work with. Compared to typical bulk RNA-seq data, the number of samples in a scRNAseq data is typically orders of magnitude higher. If differential expression (DE) between two populations of cells is of interest, and a gene-specific "count-depth relationship" confounds the DE, one may argue that we no longer need normalization before analysis. One could choose to adjust for this confounding in the regression setting, as is done in MAST (Finak et al., 2015). In a regression with sample size over several hundred, adding the library size as a covariate simply means using one degree of freedom to account for the "count-depth relationship." Since the regression is done for each gene, this allows gene specific adjustment. The drawback is that this assumes a linear effect of the library size, which may not be valid in all cells, and it can be sensitive to which cells are included in the analysis. This is also limited to the DE analysis, whereas scRNA-seq is used for many more applications.

This paper addresses normalization for scRNA-seq data in relatively high library size, without the use of unique molecular identifiers (UMI). When UMIs are used, the amplification bias is largely eliminated because multiple amplified copies of the same transcript is only counted once. These data sets still have a need for normalization because library size remains an obvious factor in the observed counts. But it is a different problem and beyond the scope of this manuscript.

4. METHODS

4.1. Probability Model

We consider each gene in any given cell is either actively transcribed or not expressed. When it is transcribed (we refer to this as Phase II or the active phase), its expression level is represented as a concentration θ_{gi} for gene g in cell i. When it is not transcribed (we referred to this as the background phase), its count depends on a sample(cell)-specific noise distribution. As described in Wu et al. (2018), we model a gene's true expected concentration as a lognormal random variable, and the background noise as a zero-inflated Poisson (ZIP) distribution. The sequencing technology does not directly measure θ_{gi}, because the RNA molecules in the cells have to be captured, reversed transcribed, amplified and eventually counted. To account for the potentially unequal counting efficiency for the RNAs of different genes in different cells, we use S_{gi} to represent the technical distortion for gene g in cell i.

The observed count thus comes from a mixture distribution with latent phase Z_{gi}, where $Z_{gi} = 1$ means the gene is in the active phase. Thus, we have

$$Y_{gi}|Z_{gi} = 1, \theta_{gi} \sim Poisson(\theta_{gi}S_{gi}) with \theta_{gi} \sim LN(\mu_g, \sigma_g^2),$$
$$Y_{gi}|Z_{gi} = 0 \sim ZIP(p_{0i}, \lambda_i)$$

The parameters θ_{gi} and S_{gi} cannot be both uniquely identified. For identifiability we constraint the average of S_{gi} for the cell with the median sequencing depth to be 1. In **Supplementary Figure 4** we show the observed log counts for a few example genes in the T2D data to illustrate that the normal assumption is a reasonable one for the active phase.

4.2. Estimating the Parameters

In Wu et al. (2018) we provide the details of the estimating procedures for obtaining the $\hat{\mu}_g, \hat{\sigma}_g^2$ and $\hat{p}_0, \hat{\lambda}$. We describe it briefly here. The ZIP parameters are estimated based on the properly of a linear relationship in the log frequency of Poisson counts, with the slope dependent on $\hat{\lambda}$. Thus, we can view the distribution of counts as ZIP contaminated by Phase II observations. We use a robust regression to down-weight the influence of high counts to obtain a robust estimate of λ and then use the amount of excessive zero to estimate p_0. The initial phase indicators Z_{gi} are set based on the point mass from the ZIP model for each observation. The parameters μ_g and σ_g are then estimated using the counts in the active phase for each gene. This is iterated using the EM algorithm, which allows us to obtain a $\hat{Z}_{gi}$ for each gene in each cell as well as $\hat{\mu}_g$.

4.3. Estimating the Normalization Factor

With these parameters we obtain residuals $\hat{\epsilon}_{gi} = \log Y_{gi} - \hat{\mu}_g$ for the genes deemed in the active phase (we use $\hat{Z}_{gi} > 0.99$), which has expectation $\log S_{gi}$ for each gene. **Figure 3A** shows an example of the distribution of the residuals against $\hat{\mu}_g$. When there is no need for normalization, $\hat{\epsilon}_{gi}$ shall be symmetrically distributed around the $y = 0$ line. When there is a consistent bias for all genes in the same cell, $\log S_{gi} \equiv \log S_i$, $\hat{\epsilon}_{gi}$ may have a non-zero expectation but will show a common trend for all expression levels. However, in general, the bias is often related to the mean expression level, as shown in **Figure 3A**. We use a spline function to estimate a smooth relationship between S_{gi} and μ_g, and obtain $\hat{f}_i$. This allows us to address the unequal need for normalization for different genes without having to put them in discrete categories. Then given a gene we estimate $\log S_{gi} = \hat{f}_i(\log Y_{gi})$.

A critical step here is to identify the genes in the active phase in a cell, as only these genes reflect the technical biases in mRNA extraction and amplification. Thus, in **Figure 3A** the smooth line is estimated using only the active phase genes (blue) only. Note that what we need is a good estimate for this curve, and thousands of genes in the active phase jointly determine

this curve. Therefore, even if for any specific gene the phase determination may not be accurate, its influence on the curve is trivial.

4.4. Use of the Normalization Factor

The normalization factor has the interpretation of the potential detection bias for gene *g* in cell *i* if gene *g* is in the active phase. This value is irrelevant in the case that the gene is not active in a cell. Directly adjusting the raw counts indiscriminately, such as in TPM, often leads to inflation of gene counts in cells with low total counts, which may create misleading large fold changes across cells. Thus, we provide the normalization factor as an offset that can be incorporated into analysis pipelines that use the count data directly. To use the normalization factor for direct adjustment, we recommend filtering genes to focus on the ones that are actively expressed.

DATA AVAILABILITY STATEMENT

The datasets used for this study can be found in the Gene Expression Omnibus (GEO) under accession numbers GSE86473, GSE85917, GSE45719, and GSE75748. The method is implemented in the R package SC2P and available at https://github.com/haowulab/SC2P.

AUTHOR CONTRIBUTIONS

ZW conceived the method. HW contributed in the development, implementation, and evaluation. KS conducted the assessment of clustering analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by the NIH award R01GM122083 and R01GM124061 for HW, and by R01GM122083 and P20GM109035 for ZW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.612670/full#supplementary-material

REFERENCES

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). Scnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14:584. doi: 10.1038/nmeth.4263

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17, 1–20. doi: 10.1186/s13059-016-1033-x

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., et al. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31, 2225–2227. doi: 10.1093/bioinformatics/btv122

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 1–13. doi: 10.1186/s13059-015-0844-5

Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13, 204–216. doi: 10.1093/biostatistics/kxr054

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi: 10.1007/BF01908075

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). Sc3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., et al. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Res.* 27, 208–222. doi: 10.1101/gr.212720.116

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, 1–9. doi: 10.1186/gb-2010-11-3-r25

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14, 618–630. doi: 10.1038/nrg3542

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Wan, S., Kim, J., and Won, K. J. (2020). Sharp: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. *Genome Res.* 30, 205–213. doi: 10.1101/gr.254557.119

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14, 414–416. doi: 10.1038/nmeth.4207

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11:41. doi: 10.1038/nmeth.2694

Wu, Z., Zhang, Y., Stitzel, M. L., and Wu, H. (2018). Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics* 34, 3340–3348. doi: 10.1093/bioinformatics/bty329

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 14 June 2021
doi: 10.3389/fgene.2021.666244

Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies

Susana Martínez Arbas[1], Susheel Bhanu Busi[1], Pedro Queirós[1], Laura de Nies[1], Malte Herold[2], Patrick May[1], Paul Wilmes[1,3], Emilie E. L. Muller[4] and Shaman Narayanasamy[1]*

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [2]Department of Environmental Research and Innovation, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, [3]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [4]Université de Strasbourg, UMR 7156 CNRS, Génétique Moléculaire, Génomique, Microbiologie, Strasbourg, France

OPEN ACCESS

Edited by:
Himel Mallick, Merck, United States

Reviewed by:
Cecilia Noecker, University of California, San Francisco, United States
Siyuan Ma, University of Pennsylvania, United States

***Correspondence:**
Susana Martínez Arbas
susana.martinez@uni.lu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 February 2021*
Accepted: *30 April 2021*
Published: *14 June 2021*

Citation:
Martínez Arbas S, Busi SB, Queirós P, de Nies L, Herold M, May P, Wilmes P, Muller EEL and Narayanasamy S (2021) Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies. Front. Genet. 12:666244. doi: 10.3389/fgene.2021.666244

In recent years, multi-omic studies have enabled resolving community structure and interrogating community function of microbial communities. Simultaneous generation of metagenomic, metatranscriptomic, metaproteomic, and (meta) metabolomic data is more feasible than ever before, thus enabling in-depth assessment of community structure, function, and phenotype, thus resulting in a multitude of multi-omic microbiome datasets and the development of innovative methods to integrate and interrogate those multi-omic datasets. Specifically, the application of reference-independent approaches provides opportunities in identifying novel organisms and functions. At present, most of these large-scale multi-omic datasets stem from spatial sampling (e.g., water/soil microbiomes at several depths, microbiomes in/on different parts of the human anatomy) or case-control studies (e.g., cohorts of human microbiomes). We believe that longitudinal multi-omic microbiome datasets are the logical next step in microbiome studies due to their characteristic advantages in providing a better understanding of community dynamics, including: observation of trends, inference of causality, and ultimately, prediction of community behavior. Furthermore, the acquisition of complementary host-derived omics, environmental measurements, and suitable metadata will further enhance the aforementioned advantages of longitudinal data, which will serve as the basis to resolve drivers of community structure and function to understand the biotic and abiotic factors governing communities and specific populations. Carefully setup future experiments hold great potential to further unveil ecological mechanisms to evolution, microbe-microbe interactions, or microbe-host interactions. In this article, we discuss the challenges, emerging strategies, and best-practices applicable to longitudinal microbiome studies ranging from sampling, biomolecular extraction, systematic multi-omic measurements, reference-independent data integration, modeling, and validation.

Keywords: microbiome, metatranscriptomics, metaproteomics, time-series, metagenomics, metabolomics, *de novo* assembly

INTRODUCTION

Advances in the study of microbial communities have highlighted their important role in natural processes, including those considered as ecosystem services for humankind (Bodelier, 2011). Complex dynamics in microbiomes at the level of composition and structure, as well as function (Heintz-Buschart and Wilmes, 2018) stem from constant adaptation of a given community toward fluctuations of abiotic and biotic factors. However, the fate of these microbial consortia in the face of perturbations is often not understood nor predictable (Muller, 2019). Longitudinal approaches are necessary to understand microbial community dynamics, as they may offer valuable insights into temporal trends and consequences of environmental forcings, when used in tandem with host-derived (Heintz-Buschart et al., 2016; Lloyd-Price et al., 2019; Mars et al., 2020) or environmental (Law et al., 2016; Herold et al., 2020) data. Longitudinal studies can be conducted using diachronic or synchronic approaches (Costa Junior et al., 2013). Herein, we discuss the capacity of longitudinal diachronic approaches as a critical tool toward studying microbial communities. We will further focus on multi-omics longitudinal studies, which leverage the power of the entire high-throughput meta-omic spectrum, namely meta-genomics (MG), -transcriptomics (MT), -proteomics (MP), and -metabolomics (MM), as they are now more feasible and affordable than ever before (Narayanasamy et al., 2015).

Overall, longitudinal multi-omics will enhance our understanding of microbial community dynamics, which could potentially bring about positive outcomes in biomedicine, biotechnology, and for the environment. However, various aspects must be considered when conducting longitudinal multi-omic microbiome studies, ranging from experimental design, bioinformatic processing, modeling, and validation. In this article, we explore challenges, considerations, and potential solutions for such studies, based on recent advances and reports (Law et al., 2016; Lloyd-Price et al., 2019; Herold et al., 2020; Martínez Arbas et al., 2021), which are applicable to both microbe-centric (e.g., soil, water) or host-centric (e.g., human gut) systems. Finally, although this article focuses on specifically longitudinal multi-omic microbiome studies, the content is generally applicable to any large-scale microbiome studies.

MULTI-OMIC CONSIDERATIONS AND EXPERIMENTAL DESIGN FOR LONGITUDINAL STUDIES

Integration of multi-omic microbiome datasets has been routinely performed, with notable instances, including studies on type-1 diabetes (Heintz-Buschart et al., 2016), cancer (Kaysen et al., 2017), healthy human gut (Tanca et al., 2017), Crohn's disease (Erickson et al., 2012), and activated sludge (Muller et al., 2014; Roume et al., 2015; Yu et al., 2019). These studies clearly demonstrate the maturity of the current microbiome multi-omics toolbox. Despite this, and to the best of our knowledge, equivalent multi-omic surveys based on extensive longitudinal microbiome sampling remain rather limited. **Table 1** lists several relevant studies of longitudinal (at least six timepoints) and multi-omic (at least two omic levels, excluding 16S amplicon sequencing) microbiome datasets.

The famous adage "*absence of evidence is not evidence of absence*" (Altman and Bland, 1995) could likely be a prelude to most microbiome studies. Hence, we discuss these studies in the context of reference-independent bioinformatics

TABLE 1 | Longitudinal multi-omic microbiome datasets and studies.

System	Sample type	Duration*	Frequency*	Total of samples	MG	MT	MP	MM	Complementary data	Studies
Human gut microbiome	Stool samples from 132 humans; healthy or with Crohn's disease or ulcerative colitis	1 year	Bi-weekly	2,965	x	x	x	x	Host genomics, transcriptomics bisulfite sequencing, serologic profiles, diet surveys, and fecal calprotectin	Lloyd-Price et al., 2019 Ruiz-Perez et al., 2021
	Stool samples of 77 individuals	6 months	Monthly	474	x			x	Host transcriptome, metabolome, cytokines, methylome, dietary survey, and physiology	Blasche et al., 2021
Activated sludge	Floating sludge islets from a single anoxic tank	1.5 year	Weekly	53	x	x	x	x	Temperature, pH, oxygen concentration, conductivity, inflow, nitrate concentration, and extracellular metabolites	Herold et al., 2020 Martínez Arbas et al., 2021
	Full- and lab-scale activated sludge	2.5 months	Weekly	10	x	x			Temperature, pH, redox potential and dissolved oxygen	Law et al., 2016

Longitudinal multi-omic data must be of least six timepoints and at least two meta-omic readouts excluding 16S amplicon sequencing. Omics data derived from host(s) are considered separate from the microbial meta-omic spectra.

**Approximate values.*

approaches, centered around *de novo* assemblies of sequencing data (MG and MT), subsequently complemented by additional omics (MP and MM, depending on their availability; **Figure 1**). Reference-independent approaches offer asymmetric advantages and opportunities in discovering novel microbial taxa and/or functionalities (Celaj et al., 2014; Narayanasamy et al., 2015; Lapidus and Korobeynikov, 2021), compared to reference-dependent methodologies (Sunagawa et al., 2013; Treangen et al., 2013). Moreover, the integration of multi-omics has been shown to yield superior output compared to single omic studies. For instance, the co-assembly of MG and MT sequencing reads was shown to improve the quality of assembled contigs (Narayanasamy et al., 2016), which in turn improves taxonomic annotation, gene calling/annotation, binning, metabolic pathway (re) construction (Muller et al., 2018; Zhou et al., 2020; Zimmermann et al., 2021), and quantification of features, e.g., taxa/genes (Narayanasamy et al., 2016). Similarly, MP spectra searches are more effective when performed against gene databases derived from MG assemblies of the same sample/environment, compared to generic databases, thus improving the recruitment of measured peptides (Tanca et al., 2016; Heyer et al., 2017; Timmins-Schiffman et al., 2017). Moreover, such a reference-independent approach may be necessary for microbial communities that are not well characterized and lack extensive unified genome or gene catalogues, such as those available for the human gut microbiome (Li et al., 2014; Almeida et al., 2021). However, most microbial communities are heterogeneous, which further complicates downstream multi-omic data processing, integration, curation, transformation, and modeling (Jiang et al., 2019). Therefore, the adherence toward standards

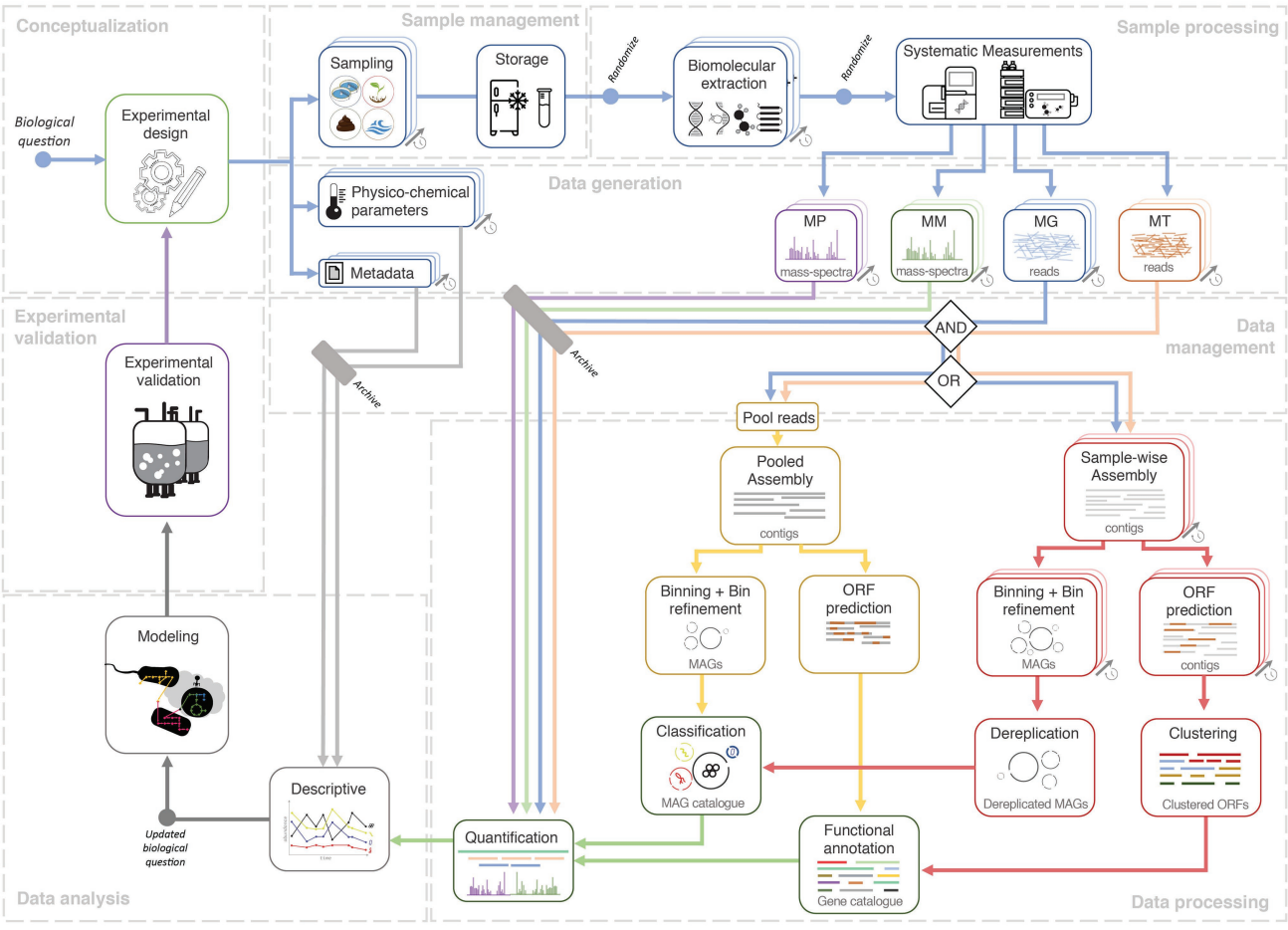


FIGURE 1 | Systems ecology workflow for longitudinal multi-omic microbiome studies. A study conceptualized *via* an experimental design phase and an initial biological question which is then followed by sample collection, sample management, and systematic high-throughput measurements. The next-generation sequencing (NGS) data could either undergo aggregated processing (yellow track) involving a pooled *de novo* assembly of NGS reads from all longitudinal samples, to eventually yield a metagenome assembled genome (MAG) and/or gene catalogue *via* binning and gene calling, respectively. In the dereplication approach (red track), data from each sample are first processed in a sample-wise manner, namely the steps of *de novo* assembly, binning, and gene calling. The resulting MAGs and predicted ORFs are then merged through a process called dereplication which generates the catalogue. The availability of a catalogue allows quantification whereby the output could be used for descriptive analyses which could potentially lead to updated or entirely novel biological questions. Quantified values, combined with descriptive analyses, could then be used within dynamic or metabolic models (gray track). Validation of models could lead to further *in situ* longitudinal experimental designs. Finally, all data (raw input, output, metadata) and code (not depicted) should be archived under a data and code management strategy. Free icons were used from https://www.flaticon.com (creators: Freepik, Gregor Cresnar, Freepik, and Smashicons).

and best-practices, spanning from sampling to data analyses is important to the outcome of a project. Accordingly, **Figure 1** illustrates the potential lifecycle of a longitudinal multi-omic microbiome study.

Longitudinal multi-omic studies require systematic and thorough study designs that consider sampling parameters (Gerber, 2014; Cao et al., 2017; Liang et al., 2020), metadata, and complementary measurements, such as physico-chemical parameters or questionnaires (Kumar et al., 2014), all of which affect downstream analyses. Sampling parameters, such as duration and frequency, are dictated by the inherent properties of a given microbial system. For instance, the sampling duration when studying gut microbiome development of neonates could span from birth until a "mature" gut microbiome composition is achieved (Stewart et al., 2018), which may vary from subject to subject. Naturally-occurring microbial systems that are exposed to the environment may exhibit annual cyclical behavior based on seasonality and, therefore, could be sampled for at least one complete season-to-season cycle (Johnston et al., 2019). Sampling frequency may be determined by the dynamics and/or generational-timescale of a given system. For instance, the human gut microbiome is known to exhibit daily fluctuations, and therefore could be sampled on a daily basis within a given temporal study (David et al., 2014), while activated sludge systems are known to exhibit (approximately) weekly doubling periods and thus could be sampled on a weekly basis (Herold et al., 2020; Martínez Arbas et al., 2021). Based on the recommendations of Sefer et al. (2016), if biological replicates are either not feasible (i.e., $n = 1$) or limited (i.e., low n) (Herold et al., 2020), one should ideally opt for higher frequency (dense) longitudinal sampling, and less dense sampling if biological replicates were available (i.e., high n), e.g., a cohort of patients (Lloyd-Price et al., 2019). Equidistant sampling is required by many downstream mathematical frameworks, such as cross-correlation or local similarity analysis (Faust et al., 2015), and thus should be strived for, as much as possible. However, the datasets listed in **Table 1**, albeit extensive and resource intensive, are not perfectly equidistant, further highlighting the practical challenges for longitudinal sampling *in situ*, including, but not limited to, accessibility, consistent biomass availability, and cost.

SAMPLE, DATA AND CODE MANAGEMENT

It is crucial to limit potential biases linked to longitudinal data, e.g., in extended time-series; samples are stored for long periods, while multiple personnel may be involved in sample collection, handling, storage, and documentation. Hence, clear guidelines and standardization must be established, as they are key factors that potentially affect downstream processes and overall outcome (Blekhman et al., 2016; Schoenenberger et al., 2016).

Biomolecular extraction from a single sample is ideal over multiple extractions from subsamples (Roume et al., 2013a). Advantageously, commercial kits for concomitant extraction of multiple biomolecules are available, including reports proposing adapted methods for extracting various biomolecules, such as DNA, total RNA, small RNA, protein, and metabolites (Peña-Llopis and Brugarolas, 2013; Roume et al., 2013b; Thorn et al., 2019). The availability of sufficient biomass (Eisenhofer et al., 2019) lysis-, homogenization-(Machiels et al., 2000; Santiago et al., 2014; Fiedorová et al., 2019) and preservation- (Borén, 2015; Hickl et al., 2019) methods are key factors that determine effectiveness to comprehensively recover all intracellular and/or extracellular biomolecules. Next, biomolecular extraction should be automated, whenever possible. While evaluations have shown that it may not necessarily provide better quality results compared to a human operator (Phillips et al., 2012), the output is more consistent (Fidler et al., 2020). In the same vein, omic readouts should also be generated on a single platform (s) as unique batches to ensure consistent output quality.

Batch effects are often overlooked in omic studies (de Goffau et al., 2021), but can be minimized during stages of sample processing by including randomization, sample tracking, and extensive documentation (Leek et al., 2010). Sample randomization implemented within batches of biomolecular extraction and high-throughput measurements could help discriminate batch effects and temporal variation, i.e., different sets of randomly selected samples from different timepoints could be treated together at each different step (Oh et al., 2019). Additionally, batch effects could be mitigated using downstream analytical (Wang and Cao, 2019) and computational methods (Gibbons et al., 2018; McLaren et al., 2019).

A potential effective experimental measure for minimizing and elucidating batch effects is the inclusion of mock/control samples during both the extraction and high-throughput measurements (Bokulich et al., 2016; Hornung et al., 2019; ATCC Mock Microbial Communities, 2020). Samples with low biomass, e.g., from neonates, glacier-streams, or acid-mine drainage, should include extraction blanks as negative controls, which are extremely valuable to discriminate contaminants arising from kits and reagents (Salter et al., 2014; Heintz-Buschart et al., 2018; Wampach et al., 2018; Weyrich et al., 2019). Furthermore, spike-ins could be helpful for downstream quantification (Zinter et al., 2019). Importantly, replicates can be used within downstream statistical frameworks (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021) to understand both within- and between-sample heterogeneity, thereby minimizing mischaracterisation of contaminants or findings driven by batch effects (de Goffau et al., 2021).

Longitudinal and multi-omic studies yield large datasets, where data processing and analyses are typically time and resource intensive. These rich datasets may be reused to study multiple aspects of a given microbial system (**Table 1**). Therefore, equal emphasis should be placed on designing bioinformatic workflows and code/data management strategies to improve reproducibility and transparency. For example, peer-review journals have begun mandating "data availability" sections and links to code repositories in adherence to project/coding best practices and standards (Sandve et al., 2013; Bokulich et al., 2020), further improving posterior data integration and analysis in the short-term, while improving scaling-up

and knowledge transfer in the long run (Shahin et al., 2017; Wilson et al., 2017). In addition, format-free archival repositories, such as Zenodo could be used for non-standard data types,[1] for instance simulated raw data, physico-chemical measurements, intermediate data, large tables, and archived Github repositories. Despite this, reports indicate that 26% of bioinformatics tools are no longer available (Mangul et al., 2019), while gaps in available raw data (Jurburg et al., 2020) and metadata (Schriml et al., 2020) still exist.

CONSTRUCTION OF LONGITUDINAL GENE AND GENOME REFERENCE CATALOGUES

Microbiomes may be studied from a gene-centric perspective (Roume et al., 2015), which requires read or contig-level taxonomic classification (Segata et al., 2012; Wood and Salzberg, 2014), ORF prediction (Hyatt et al., 2010; Rho et al., 2010), and gene annotation (Seemann, 2014; Buchfink et al., 2015; Franzosa et al., 2018; Queirós et al., 2020). Metagenome assembled genomes (MAGs) provide genomic context and can be obtained through binning (Chen et al., 2020; Yue et al., 2020) followed by taxonomic classification (Bremges et al., 2020; Chaumeil et al., 2020) and functional annotation. In that regard, several tools exist that improve the binning process by automating the selection of highest-quality MAGs (bins) and/or performing MAG refinement (Broeksema et al., 2017; Sieber et al., 2018; Uritskiy et al., 2018). These tools enable ensemble binning approaches, balancing out the strengths and weaknesses of different binning methods (Chen et al., 2020; Yue et al., 2020).

Features (i.e., taxa or genes) appear in varying quantities, in different timepoints of longitudinal meta-omic studies. It is challenging to link and track features from one timepoint to another without any given point of reference. Therefore, the construction of what we term as "representative longitudinal catalogues" (hereafter referred to as catalogues) of MAGs/genes, provides a non-redundant representative base to link features from the different longitudinal samples (Herold et al., 2020; Martínez Arbas et al., 2021). The outcome of any downstream analysis is highly reliant on the quality of the MAGs and genes within a catalogue, which further depends on the quality of large-scale bioinformatic processing (e.g., *de novo* assembly and binning). **Figure 1** illustrates two methods of constructing such catalogues, which are through aggregated processing of data from all samples or through de-replicating the output from individually processed sample data (i.e., sample-wise processing). A third alternative to these methods could be the representation of non-redundant genes in pangenomes from MAGs annotated at the species-level (Tettelin et al., 2005; Delmont and Eren, 2018), collected across all timepoints. This allows for identifying any varying patterns especially in the context of environmental factors and phylogenetic constraints influencing gene acquisition and/or genome-streamlining (Tettelin et al., 2005). Given that others have highlighted the catalogue building methodologies (Qin et al., 2010; Nayfach et al., 2020; Almeida et al., 2021); here, we elaborate methods discussed above in the context of both gene- and MAG-centric strategies.

The general advantage of the aggregated processing approach is simplicity, whereby a single run is required for all the large-scale bioinformatic processing steps (**Figure 1**). Moreover, pooled assemblies have been shown to be effective (Magasin and Gerloff, 2015), especially in the advent of highly efficient *de novo* assemblers (Li et al., 2016) and digital normalization (Brown et al., 2012). However, pooling reads from a large number of samples increases the complexity of the *de novo* assembly process, especially for complex communities. It also requires substantial computational resources, while potentially resulting in lower quality contigs, MAGs, and genes (Chen et al., 2020).

The dereplication method (**Figure 1**) is applied after independent sample-wise large-scale bioinformatic processing (Evans and Denef, 2020). Predicted ORFs could be de-replicated through clustering (Li and Godzik, 2006; Edgar, 2010; Mirdita et al., 2019), producing a gene catalogue (Li et al., 2014). On the contrary, the dereplication of MAGs is more complex, requiring several steps: binning from sample-wise *de novo* assemblies to generate MAGs, curation of high-quality MAGs (Parks et al., 2015), and dereplication of MAGs (Olm et al., 2017; Wampach et al., 2018) to select the most representative MAGs of the longitudinal data (Uritskiy et al., 2018; Chen et al., 2020). In general, dereplication methods are particularly advantageous for longitudinal microbiome studies with many deeply sequenced samples (Herold et al., 2020; Martínez Arbas et al., 2021).

Although not systematically evaluated, one caveat worth considering when constructing a catalogue based on *de novo* assemblies, binning, and dereplication is the potential loss of resolution in population-level diversity (Kashtan et al., 2014; Evans and Denef, 2020; Quince et al., 2020), which may include single nucleotide variants, copy number variants, strains, and auxiliary gene content (Evans and Denef, 2020) potentially impacting important downstream steps, such as integration of metaproteomic data (Tanca et al., 2016) or time-resolved strain tracking (Brito and Alm, 2016; Zlitni et al., 2020). To the best of our knowledge, the extent of the impact has yet to be systematically investigated. In our opinion, several strategies can be applied to overcome this issue, including the usage of a comparative genomics methodology, i.e., pangenomes (Delmont and Eren, 2018), even opt for (re) assemblies of read subsets associated to particular taxa or MAGs of interest (Albertsen et al., 2013), or the application of strain-level analysis tools (Anyansi et al., 2020).

Overall, choosing the specific methods for constructing a longitudinal catalogue depends on various factors, including the biological question, complexity of the community (van der Walt et al., 2017), number of samples, and sequencing depth. To the best of our knowledge, a comparison between an aggregated processing approach and a dereplication approach has yet to be conducted. Such a comparison would further help to inform researchers on selecting the best strategy for longitudinal analyses.

[1]https://zenodo.org

QUANTIFICATION AND NORMALIZATION

Longitudinal catalogues provide compositional information of community taxa and potential functions. However, the relative quantification of community members and functionalities is key in harnessing the power of longitudinal microbiome data, as it allows the observation of community taxa/functional dynamics and could be used in downstream modeling. In that regard, quantifying MG and MT sequencing data is a standard process of aligning reads (Li and Durbin, 2009) to relevant catalogues, and then quantifying features of interest (e.g., population/gene relative genomic abundance, gene expression) based on those alignments, providing information on community structure, functional potential, and gene expression. Complementally, MP data provide functional insights, whereby several methods are available for the quantification of such data (Delogu et al., 2020; Pible et al., 2020), while identification and quantification of metabolites through MM data (Kapoore and Vaidyanathan, 2016; Mallick et al., 2019; Røst et al., 2020) provide insights on the community phenotype (s). However, *in situ* measurements of substrate uptake through labeling-based approaches (Starr et al., 2018) are challenging. Therefore, specific metabolites of interest could be indirectly linked to members of a microbial community by proportionally assigning the relative contribution of a MAG to a given (re) constructed metabolic pathway based on genomic abundance or gene/protein expression (Noecker et al., 2016; Blasche et al., 2021).

Normalization of quantified values is required to enable community structure and function comparisons between timepoint samples. The selection of normalization methods is important as it affects downstream analytical steps. There are several methods to normalize longitudinal MG and MT data, from the generation of compositional data to log-ratios and differential rankings (Chen et al., 2018; Pereira et al., 2018; Morton et al., 2019). Additionally, one should also inspect the data for potential confounding batch effects and take it into consideration when performing normalization (Gibbons et al., 2018; McLaren et al., 2019; Coenen et al., 2020). In summary, effective relative quantification and normalization will serve as a strong basis for downstream modeling approaches, and the development of robust methods for absolute quantification will be decisive in the future.

ANALYSIS OF COMMUNITY CHARACTERISTICS AND DYNAMICS

Generally, microbiome omic data are complex, as it is (i) compositional, e.g., provided as relative abundances, which require specific considerations when selecting statistical analyses (Gloor et al., 2017), (ii) highly sparse, such that the interpretation of zero-values generated from sampling, biological, or technical processes heavily affects data-derived conclusions (Silverman et al., 2020), and (iii) high dimensional, which increases modeling difficulty due to the influence of feature selection that heavily affect potential predictions (Bolón-Canedo et al., 2016). Furthermore, multi-omic studies may contain gaps within the omic spectrum, such that certain samples may not be represented within a certain omic layer (Lloyd-Price et al., 2019). Despite introducing complexity, the complementary use of different omics could improve analysis outcomes and add predictive power to models (Muller et al., 2013; Fondi and Liò, 2015). Longitudinal data introduce another layer of complexity, i.e., time dependencies, such that one timepoint is dependent on the previous timepoints, rendering conventional statistical analyses unsuitable as they assume samples to be independent (Coenen et al., 2020). This is further compounded by the fact that samples from longitudinal *in situ* studies are often low in number and non-equidistant (Park et al., 2020). Imputation may be used to supplement missing values (i.e., omic measurements or timepoints; Jiang et al., 2020).

Initial exploration of the microbiome dynamics can be assessed through ordination analyses, where high dimensional population structure data are visualized in a two-dimensional space to observe the trajectory of the samples and the behavior of the system, i.e., metastability, cycles, and alternative states (Gonze et al., 2018). Then, community member relationships may be inferred using, e.g., correlation methods (Faust et al., 2012; Friedman and Alm, 2012; Weiss et al., 2016). Unfortunately, correlations may be insufficient to assess complex community interactions, whereby the application of modeling approaches would be necessary to resolve those relationships (Fisher and Mehta, 2014; Trosvik et al., 2015; Ridenhour et al., 2017). Modeling could serve as a means of integrating several layers of omic data (Lloyd-Price et al., 2019; Ruiz-Perez et al., 2021) further elucidating microbial interplay beyond species abundances and functional potential.

Extensive literature of statistical and mathematical frameworks for multi-omic and/or longitudinal microbiome data is currently available. For instance, Noor et al. (2019) review the integration of multi-omics data from data-driven and knowledge-based perspectives. Coenen et al. (2020) discuss approaches to characterize temporal dynamics and to identify periodicity of populations and putative interactions between them, while Faust et al. (2018) propose a classification scheme for better model selection. Bodein et al. (2019) provide a multivariate framework to integrate longitudinal and multi-omics data, while Park et al. (2020) discuss the development of models and software tools for time-series metagenome and metabolome data. Overall, the application of these methodologies should be tailored toward specific hypotheses and studies, for which data exploration is essential to select modeling approaches that fit the type, quality, and quantity of the data.

More recently, the emergence of studies which track microbiome dynamics of cohorts over time, i.e., multiple individuals/sites (Carmody et al., 2019; Lloyd-Price et al., 2019; Mars et al., 2020), necessitates the ability to discriminate variation stemming from the same individual/environment compared to those from different individuals/environments. In such cases, multi-level statistical modeling (also known as mixed-effects/hierarchical models) is able to account for repeated sampling or nested variation across a sample population (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021). Most notably Lloyd-Price et al. (2019) extensively applied such

methods to associate multi-omic microbiome signatures with host-derived molecular profiles in a cohort of 132 individuals. Other instances include multi-omic longitudinal studies that combine murine and human datasets to unveil the adaptation of gut microbiomes to raw and cooked food (Carmody et al., 2019) and the identification of therapeutic targets for irritable bowel syndrome (Mars et al., 2020). Finally, there are newer methodologies that apply similar/related statistical frameworks to modeling multi-omic data (Mallick et al., 2021).

The validation of the models remains one of the most challenging issues. Mathematical models combined with culture of synthetic microbial communities are commonly utilized to study mechanisms behind host-microbiome interactions (Moejes et al., 2017). It is also possible to validate interactions between microbes by, e.g., applying environmental perturbations in controlled conditions (Law et al., 2016; Herold et al., 2020). These explorations may result in a further understanding of the role of biotic and abiotic factors in shaping microbiomes, in relation to community phenotypes found in nature, biotechnological processes (Law et al., 2016; Herold et al., 2020), or host-associated microbiomes (Moejes et al., 2017; Garza et al., 2018).

CONCLUSION

Longitudinal microbiome studies combined with integrated multi-omic measurements provide unprecedented opportunities to study microbial community dynamics, both structurally and functionally. In tandem with evolving high-throughput technologies, e.g., long-read sequencing (Moss et al., 2020; Wickramarachchi et al., 2020), these studies will become important tools in the exploration and potential exploitation of microbial consortia. We described strategies to mitigate the various challenges associated with such studies, encompassing study design, best practices, practical considerations, and bioinformatics processing and modeling. While longitudinal multi-omics datasets are currently scarce (**Table 1**), we are confident that it will increasingly become more common, similar to how we are increasingly transitioning from single omics to multi-omic (Noor et al., 2019). Longitudinal microbiome multi-omics will serve as an important tool for further improving analytical methods, which will in turn lead to relevant biomedical, biotechnological, and environmental outcomes.

AUTHOR CONTRIBUTIONS

SMA and SN outlined the manuscript and coordinated the writing process. LdN, SN, and SMA prepared the figure. All authors contributed to the writing, reviewing, and editing of the manuscript. All authors approved the submitted version.

FUNDING

The Luxembourg National Research Fund (FNR) supported SMA, PQ, LdN, PM, and EELM through the PRIDE doctoral training unit grants (PRIDE15/10907093) and (PRIDE/18/11823097), the CORE Junior grant (C15/SR/10404839), and the CORE grant (CORE/17/SM/11689322). SBB was supported by the Sinergia grant (CRSII5_180241) through the Swiss National Science Foundation. PW was supported by the European Research Council (ERC-CoG 863664).

ACKNOWLEDGMENTS

We would like to thank Oskar Hickl for his input on metaproteomic analysis.

REFERENCES

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., et al. (2021). A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114. doi: 10.1038/s41587-020-0603-3

Altman, D. G., and Bland, J. M. (1995). Statistics notes: absence of evidence is not evidence of absence. *BMJ* 311:485. doi: 10.1136/bmj.311.7003.485

Anderson, M. J. (2017). "Permutational multivariate analysis of variance (PERMANOVA)," in *Wiley Stats Ref: Statistics Reference Online*. eds. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels (Chichester, UK: John Wiley & Sons, Ltd.), 1–15.

Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., and Abeel, T. (2020). Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Front. Microbiol.* 11:1925. doi: 10.3389/fmicb.2020.01925

ATCC Mock Microbial Communities (2020). Available at: https://www.atcc.org/en/Products/Microbiome_Standards.aspx (Accessed November 30, 2020).

Blasche, S., Kim, Y., Mars, R. A. T., Machado, D., Maansson, M., Kafkia, E., et al. (2021). Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nat. Microbiol.* 6, 196–208. doi: 10.1038/s41564-020-00816-5

Blekhman, R., Tang, K., Archie, E. A., Barreiro, L. B., Johnson, Z. P., Wilson, M. E., et al. (2016). Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. *Sci. Rep.* 6:31519. doi: 10.1038/srep31519

Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* 10:963. doi: 10.3389/fgene.2019.00963

Bodelier, P. L. E. (2011). Toward understanding, managing, and protecting microbial ecosystems. *Front. Microbiol.* 2:80. doi: 10.3389/fmicb.2011.00080

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., et al. (2016). Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 1:e00062-16. doi: 10.1128/mSystems.00062-16

Bokulich, N. A., Ziemski, M., Robeson, M. S., and Kaehler, B. D. (2020). Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Comput. Struct. Biotechnol. J.* 18, 4048–4062. doi: 10.1016/j.csbj.2020.11.049

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Prog. Artif. Intell.* 5, 65–75. doi: 10.1007/s13748-015-0080-y

Borén, M. (2015). "Sample preservation Through heat stabilization of proteins: principles and examples," in *Proteomic Profiling Methods in Molecular Biology*. ed. A. Posch (New York, NY: Springer), 21–32.

Bremges, A., Fritz, A., and McHardy, A. C. (2020). CAMITAX: taxon labels for microbial genomes. *Giga Science* 9:giz154. doi: 10.1093/gigascience/giz154

Brito, I. L., and Alm, E. J. (2016). Tracking strains in the microbiome: insights from metagenomics and models. *Front. Microbiol.* 7:712. doi: 10.3389/fmicb.2016.00712

Broeksema, B., Calusinska, M., McGee, F., Winter, K., Bongiovanni, F., Goux, X., et al. (2017). ICoVeR – an interactive visualization tool for verification and refinement of metagenomic bins. *BMC Bioinformatics* 18:233. doi: 10.1186/s12859-017-1653-5

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv [Preprint].

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cao, H.-T., Gibson, T. E., Bashan, A., and Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: pitfalls and lessons. *BioEssays* 39:1600188. doi: 10.1002/bies.201600188

Carmody, R. N., Bisanz, J. E., Bowen, B. P., Maurice, C. F., Lyalina, S., Louie, K. B., et al. (2019). Cooking shapes the structure and function of the gut microbiome. *Nat. Microbiol.* 4, 2052–2063. doi: 10.1038/s41564-019-0569-4

Celaj, A., Markle, J., Danska, J., and Parkinson, J. (2014). Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome* 2:39. doi: 10.1186/2049-2618-2-39

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res.* 30, 315–333. doi: 10.1101/gr.258640.119

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A primer for microbiome time-series analysis. *Front. Genet.* 11:310. doi: 10.3389/fgene.2020.00310

Costa Junior, C., Corbeels, M., Bernoux, M., Píccolo, M. C., Siqueira Neto, M., Feigl, B. J., et al. (2013). Assessing soil carbon storage rates under no-tillage: comparing the synchronic and diachronic approaches. *Soil Tillage Res.* 134, 207–212. doi: 10.1016/j.still.2013.08.010

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89

de Goffau, M. C., Charnock-Jones, D. S., Smith, G. C. S., and Parkhill, J. (2021). Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. *Microbiome* 9:6. doi: 10.1186/s40168-020-00949-z

Delmont, T. O., and Eren, A. M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* 6:e4320. doi: 10.7717/peerj.4320

Delogu, F., Kunath, B. J., Evans, P. N., Arntzen, M. Ø., Hvidsten, T. R., and Pope, P. B. (2020). Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat. Commun.* 11:4708. doi: 10.1038/s41467-020-18543-0

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* 2, 105–117. doi: 10.1016/j.tim.2018.11.003

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 7:e49138. doi: 10.1371/journal.pone.0049138

Evans, J. T., and Denef, V. J. (2020). To dereplicate or not to dereplicate? *mSphere* 5:e00971-19. doi: 10.1128/mSphere.00971-19

Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606

Fidler, G., Tolnai, E., Stagel, A., Remenyik, J., Stundl, L., Gal, F., et al. (2020). Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. *Sci. Rep.* 10:3419. doi: 10.1038/s41598-020-60304-y

Fiedorová, K., Radvanský, M., Němcová, E., Grombiříková, H., Bosák, J., Černochová, M., et al. (2019). The impact of DNA extraction methods on stool bacterial and fungal microbiota community recovery. *Front. Microbiol.* 10:821. doi: 10.3389/fmicb.2019.00821

Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 9:e102451. doi: 10.1371/journal.pone.0102451

Fondi, M., and Liò, P. (2015). Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiol. Res.* 171, 52–64. doi: 10.1016/j.micres.2015.01.003

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968. doi: 10.1038/s41592-018-0176-y

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Garza, D. R., van Verk, M. C., Huynen, M. A., and Dutilh, B. E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* 3, 456–460. doi: 10.1038/s41564-018-0124-8

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037

Gibbons, S. M., Duvallet, C., and Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.* 14:e1006102. doi: 10.1371/journal.pcbi.1006102

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gonze, D., Coyte, K. Z., Lahti, L., and Faust, K. (2018). Microbial communities as dynamical systems. *Curr. Opin. Microbiol.* 44, 41–49. doi: 10.1016/j.mib.2018.07.004

Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* 2:16180. doi: 10.1038/nmicrobiol.2016.227

Heintz-Buschart, A., and Wilmes, P. (2018). Human gut microbiome: function matters. *Trends Microbiol.* 26, 563–574. doi: 10.1016/j.tim.2017.11.002

Heintz-Buschart, A., Yusuf, D., Kaysen, A., Etheridge, A., Fritz, J. V., May, P., et al. (2018). Small RNA profiling of low biomass samples: identification and removal of contaminants. *BMC Biol.* 16:52. doi: 10.1186/s12915-018-0522-7

Herold, M., Arbas, S. M., Narayanasamy, S., Sheik, A. R., Kleine-Borgmann, L. A. K., Lebrun, L. A., et al. (2020). Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat. Commun.* 11:5281. doi: 10.1038/s41467-020-19006-2

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* 261, 24–36. doi: 10.1016/j.jbiotec.2017.06.1201

Hickl, O., Heintz-Buschart, A., Trautwein-Schult, A., Hercog, R., Bork, P., Wilmes, P., et al. (2019). Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut microbiome. *Microorganisms* 7:367. doi: 10.3390/microorganisms7090367

Hornung, B. V. H., Zwittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95:fiz045. doi: 10.1093/femsec/fiz045

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., et al. (2019). Microbiome multi-omics network analysis: statistical considerations,

limitations, and opportunities. *Front. Genet.* 10:995. doi: 10.3389/fgene.2019.00995

Jiang, R., Li, W. V., and Li, J. J. (2020). mbImpute: an accurate and robust imputation method for microbiome data. *Genomics* [Preprint]. doi: 10.1101/2020.03.07.982314

Johnston, J., LaPara, T., and Behrens, S. (2019). Composition and dynamics of the activated sludge microbiome during seasonal nitrification failure. *Sci. Rep.* 9:4565. doi: 10.1038/s41598-019-40872-4

Jurburg, S. D., Konzack, M., Eisenhauer, N., and Heintz-Buschart, A. (2020). The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Commun. Biol.* 3:474. doi: 10.1038/s42003-020-01204-9

Kapoore, R. V., and Vaidyanathan, S. (2016). Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374:20150363. doi: 10.1098/rsta.2015.0363

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344, 416–420. doi: 10.1126/science.1248575

Kaysen, A., Heintz-Buschart, A., Muller, E. E. L., Narayanasamy, S., Wampach, L., Laczny, C. C., et al. (2017). Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic hematopoietic stem cell transplantation. *Transl. Res.* 186, 79–94. doi: 10.1016/j.trsl.2017.06.008

Kumar, R., Eipers, P., Little, R. B., Crowley, M., Crossman, D. K., Lefkowitz, E. J., et al. (2014). Getting started with microbiome analysis: sample acquisition to bioinformatics. *Curr. Protoc. Hum. Genet.* 82, 18.8.1–18.8.29. doi: 10.1002/0471142905.hg1808s82

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lapidus, A. L., and Korobeynikov, A. I. (2021). Metagenomic data assembly – the way of decoding unknown microorganisms. *Front. Microbiol.* 12:613791. doi: 10.3389/fmicb.2021.613791

Law, Y., Kirkegaard, R. H., Cokro, A. A., Liu, X., Arumugam, K., Xie, C., et al. (2016). Integrative microbial community analysis reveals full-scale enhanced biological phosphorus removal under tropical conditions. *Sci. Rep.* 6:25719. doi: 10.1038/srep25719

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11. doi: 10.1016/j.ymeth.2016.02.020

Liang, Y., Dong, T., Chen, M., He, L., Wang, T., Liu, X., et al. (2020). Systematic analysis of impact of sampling regions and storage methods on Fecal gut microbiome and metabolome profiles. *mSphere* 5:e00763-19. doi: 10.1128/mSphere.00763-19

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi: 10.1038/s41586-019-1237-9

Machiels, B. M., Ruers, T., Lindhout, M., Hardy, K., Hlavaty, T., Bang, D. D., et al. (2000). New protocol for DNA extraction of stool. *Bio Techniques* 28, 286–290. doi: 10.2144/00282st05

Magasin, J. D., and Gerloff, D. L. (2015). Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics* 31, 311–317. doi: 10.1093/bioinformatics/btu546

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10:3136. doi: 10.1038/s41467-019-10927-1

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *Microbiology* [Preprint]. doi: 10.1099/mic.0.001031

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20:47. doi: 10.1186/s13059-019-1649-8

Mars, R. A. T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., et al. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* 182, 1460–1473. doi: 10.1016/j.cell.2020.08.007

Martínez Arbas, S. M., Narayanasamy, S., Herold, M., Lebrun, L. A., Hoopmann, M. R., Li, S., et al. (2021). Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nat. Microbiol.* 6, 123–135. doi: 10.1038/s41564-020-00794-8

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *elife* 8:e46923. doi: 10.7554/eLife.46923

Mirdita, M., Steinegger, M., and Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35, 2856–2858. doi: 10.1093/bioinformatics/bty1057

Moejes, F., Succurro, A., Popa, O., Maguire, J., and Ebenhöh, O. (2017). Dynamics of the bacterial community associated with *Phaeodactylum tricornutum* cultures. *Processes* 5:77. doi: 10.3390/pr5040077

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* 10:2719. doi: 10.1038/s41467-019-10656-5

Moss, E. L., Maghini, D. G., and Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* 38, 701–707. doi: 10.1038/s41587-020-0422-6

Muller, E. E. L. (2019). Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* 4:e00080-19. doi: 10.1128/mSystems.00080-19

Muller, E. E. L., Faust, K., Widder, S., Herold, M., Arbas, S. M., and Wilmes, P. (2018). Using metabolic networks to resolve ecological properties of microbiomes. *Curr. Opin. Syst. Biol.* 8, 73–80. doi: 10.1016/j.coisb.2017.12.004

Muller, E. E. L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends Microbiol.* 21, 325–333. doi: 10.1016/j.tim.2013.04.009

Muller, E. E. L., Pinel, N., Laczny, C. C., Hoopmann, M. R., Narayanasamy, S., Lebrun, L. A., et al. (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* 5:5603. doi: 10.1038/ncomms6603

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Heintz-Buschart, A., Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17:260. doi: 10.1186/s13059-016-1116-8

Narayanasamy, S., Muller, E. E. L., Sheik, A. R., and Wilmes, P. (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microb. Biotechnol.* 8, 363–368. doi: 10.1111/1751-7915.12255

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2020). A genomic catalog of earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi: 10.1038/s41587-020-0718-6

Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., et al. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 1:e00013-15. doi: 10.1128/mSystems.00013-15

Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Gene Regul.* 15, 39–47. doi: 10.1016/j.coisb.2019.03.007

Oh, S., Li, C., Baldwin, R. L., Song, S., Liu, F., and Li, R. W. (2019). Temporal dynamics in meta longitudinal RNA-Seq data. *Sci. Rep.* 9:763. doi: 10.1038/s41598-018-37397-7

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

Park, S.-Y., Ufondu, A., Lee, K., and Jayaraman, A. (2020). Emerging computational tools and models for studying gut microbiota composition and function. *Tissue Cell Pathw. Eng.* 66, 301–311. doi: 10.1016/j.copbio.2020.10.005

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Check M: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Peña-Llopis, S., and Brugarolas, J. (2013). Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications. *Nat. Protoc.* 8, 2240–2255. doi: 10.1038/nprot.2013.141

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274. doi: 10.1186/s12864-018-4637-6

Phillips, K., McCallum, N., and Welch, L. (2012). A comparison of methods for forensic DNA extraction: Chelex-100® and the QIAGEN DNA Investigator Kit (manual and automated). *Forensic Sci. Int. Genet.* 6, 282–285. doi: 10.1016/j.fsigen.2011.04.018

Pible, O., Allain, F., Jouffret, V., Culotta, K., Miotello, G., and Armengaud, J. (2020). Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome* 8:30. doi: 10.1186/s40168-020-00797-x

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Queirós, P., Delogu, F., Hickl, O., May, P., and Wilmes, P. (2020). Mantis: flexible and consensus-driven genome annotation. *Bioinformatics* [Preprint]. doi: 10.1101/2020.11.02.360933

Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O. S., Summers, J. K., et al. (2020). Metagenomics strain resolution on assembly graphs. *Bioinformatics* [Preprint]. doi: 10.1101/2020.09.06.284828

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., et al. (2017). Modeling time-series data from microbial communities. *ISME J.* 11, 2526–2537. doi: 10.1038/ismej.2017.107

Røst, L. M., Brekke Thorfinnsdottir, L., Kumar, K., Fuchino, K., Eide Langørgen, I., Bartosova, Z., et al. (2020). Absolute quantification of the central carbon metabolome in eight commonly applied prokaryotic and eukaryotic model systems. *Metabolites* 10:74. doi: 10.3390/metabo10020074

Roume, H., Heintz-Buschart, A., Muller, E. E. L., May, P., Satagopam, V. P., Laczny, C. C., et al. (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *Npj Biofilms Microbiomes* 1:15007. doi: 10.1038/npjbiofilms.2015.7

Roume, H., Heintz-Buschart, A., Muller, E. E. L., and Wilmes, P. (2013b). "Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample," in *Methods in Enzymology*. ed. E. F. DeLong (Cambridge, Massachusetts, United States: Elsevier), 219–236.

Roume, H., Muller, E. E., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013a). A biomolecular isolation framework for eco-systems biology. *ISME J.* 7, 110–121. doi: 10.1038/ismej.2012.72

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., et al. (2021). Dynamic Bayesian networks for integrating multi-omics time series microbiome data. *mSystems* 6:e01105-20. doi: 10.1128/mSystems.01105-20

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* 9:e1003285. doi: 10.1371/journal.pcbi.1003285

Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., et al. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* 14:112. doi: 10.1186/1471-2180-14-112

Schoenenberger, A. W., Muggli, F., Parati, G., Gallino, A., Ehret, G., Suter, P. M., et al. (2016). Protocol of the Swiss Longitudinal Cohort Study (SWICOS) in rural Switzerland. *BMJ Open* 6:e013280. doi: 10.1136/bmjopen-2016-013280

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., et al. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* 7:188. doi: 10.1038/s41597-020-0524-5

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sefer, E., Kleyman, M., and Bar-Joseph, Z. (2016). Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments. *Cell Syst.* 3, 35–42. doi: 10.1016/j.cels.2016.06.007

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Shahin, M., Ali Babar, M., and Zhu, L. (2017). Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* 5, 3909–3943. doi: 10.1109/ACCESS.2017.2685629

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843. doi: 10.1038/s41564-018-0171-1

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Sokal, R. R. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research. 3rd Edn.* New York: W.H. Freeman.

Starr, E. P., Shi, S., Blazewicz, S. J., Probst, A. J., Herman, D. J., Firestone, M. K., et al. (2018). Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. *Microbiome* 6:122. doi: 10.1186/s40168-018-0499-z

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588. doi: 10.1038/s41586-018-0617-x

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Tanca, A., Abbondio, M., Palomba, A., Fraumene, C., Manghina, V., Cucca, F., et al. (2017). Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* 5:79. doi: 10.1186/s40168-017-0293-3

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., et al. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* 4:51. doi: 10.1186/s40168-016-0196-8

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Thorn, C. E., Bergesch, C., Joyce, A., Sambrano, G., McDonnell, K., Brennan, F., et al. (2019). A robust, cost-effective method for DNA, RNA and protein co-extraction from soil, other complex microbiomes and pure cultures. *Mol. Ecol. Resour.* 19, 439–455. doi: 10.1111/1755-0998.12979

Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., et al. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* 11, 309–314. doi: 10.1038/ismej.2016.132

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

Trosvik, P., de Muinck, E. J., and Stenseth, N. C. (2015). Biotic interactions and temporal dynamics of the human gastrointestinal microbiota. *ISME J.* 9, 533–541. doi: 10.1038/ismej.2014.147

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1

van der Walt, A. J., van Goethem, M. W., Ramond, J.-B., Makhalanyane, T. P., Reva, O., and Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18:521. doi: 10.1186/s12864-017-3918-9

Wampach, L., Heintz-Buschart, A., Fritz, J. V., Ramiro-Garcia, J., Habier, J., Herold, M., et al. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* 9:5091. doi: 10.1038/s41467-018-07631-x

Wang, Y., and Cao, K.-A. L. (2019). Managing batch effects in microbiome data. *Brief. Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., et al. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi: 10.1111/1755-0998.13011

Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., and Lin, Y. (2020). MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics* 36, i3–i11. doi: 10.1093/bioinformatics/btaa441

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Comput. Biol.* 13:e1005510. doi: 10.1371/journal.pcbi.1005510

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Yu, K., Yi, S., Li, B., Guo, F., Peng, X., Wang, Z., et al. (2019). An integrated meta-omics approach reveals substrates involved in synergistic interactions in a bisphenol A (BPA)-degrading microbial community. *Microbiome* 7:16. doi: 10.1186/s40168-019-0634-5

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., et al. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 21:334. doi: 10.1186/s12859-020-03667-3

Zhou, Z., Tran, P. Q., Breiser, A. M., Liu, Y., Kieft, K., Cowley, E. S., et al. (2020). METABOLIC: high-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.27.357558

Zimmermann, J., Kaleta, C., and Waschina, S. (2021). gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* 22:81. doi: 10.1186/s13059-021-02295-1

Zinter, M. S., Mayday, M. Y., Ryckman, K. K., Jelliffe-Pawlowski, L. L., and DeRisi, J. L. (2019). Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* 7, 62. doi: 10.1186/s40168-019-0678-6

Zlitni, S., Bishara, A., Moss, E. L., Tkachenko, E., Kang, J. B., Culver, R. N., et al. (2020). Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Med.* 12:50. doi: 10.1186/s13073-020-00747-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 October 2021
doi: 10.3389/fgene.2021.763431

GeneMarkeR: A Database and User Interface for scRNA-seq Marker Genes

Brianna M. Paisley[1,2]* *and Yunlong Liu*[1,3]

[1]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN, United States, [2]Toxicology, Eli Lilly and Company, Indianapolis, IN, United States, [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States

Single-cell sequencing (scRNA-seq) has enabled researchers to study cellular heterogeneity. Accurate cell type identification is crucial for scRNA-seq analysis to be valid and robust. Marker genes, genes specific for one or a few cell types, can improve cell type classification; however, their specificity varies across species, samples, and cell subtypes. Current marker gene databases lack standardization, cell hierarchy consideration, sample diversity, and/or the flexibility for updates as new data become available. Most of these databases are derived from a single statistical analysis despite many such analyses scattered in the literature to identify marker genes from scRNA-seq data and pure cell populations. An R Shiny web tool called GeneMarkeR was developed for researchers to retrieve marker genes demonstrating cell type specificity across species, methodology and sample types based on a novel algorithm. The web tool facilitates online submission and interfaces with MySQL to ensure updatability. Furthermore, the tool incorporates reactive programming to enable researchers to retrieve standardized public data supporting the marker genes. GeneMarkeR currently hosts over 261,000 rows of standardized marker gene results from 25 studies across 21,012 unique genomic entities and 99 unique cell types mapped to hierarchical ontologies.

Keywords: single-cell RNA-seq1, scRNA-seq2, marker gene3, cell type4, database5, web-interface6

OPEN ACCESS

Edited by:
Himel Mallick,
Merck, United States

Reviewed by:
Haoyun Lei,
Carnegie Mellon University, United States
Jun Li,
University of Notre Dame, United States

***Correspondence:**
Brianna M. Paisley
bpaisley@iu.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 August 2021*
Accepted: *16 September 2021*
Published: *26 October 2021*

Citation:
Paisley BM and Liu Y (2021) GeneMarkeR: A Database and User Interface for scRNA-seq Marker Genes. Front. Genet. 12:763431. doi: 10.3389/fgene.2021.763431

INTRODUCTION

scRNA-seq enables study of disease heterogeneity, novel cell subtypes, cellular interactions, and cellular tissue composition (Mancarci et al., 2017; Skelly et al., 2018; Aran et al., 2019; Saviano et al., 2020). A major challenge in scRNA-seq analysis is to identify the cell type of individual cells. Accurate cell type identification is crucial for any scRNA-seq analysis to be valid as incorrect cell type assignment will reduce statistical robustness and may lead to incorrect biological conclusions. Therefore, accurate and comprehensive cell type assignment is necessary for reliable biological insights into scRNA-seq datasets.

Marker genes, genes more specific in expression for one or a few cell types over others, are important descriptors in the identification of scRNA-seq cell type (Franzen et al., 2019; Zhang et al., 2019). Identifying marker genes can be a tedious process, and sometimes requires manual extraction from appendices and/or images of publications. Furthermore, marker genes may be specific to sample type, species, and/or sequencing technology. For example, a gene that is specific for endothelial cells in mouse brain tissue samples may not be endothelial cell specific outside of the brain or in human samples. Therefore, it is vital to improve access to accurate, robust, and translatable scRNA-seq marker genes.

The recent publication of CellMarker (Zhang et al., 2019) has provided researchers with access to marker gene lists in mouse and human. The program provides manually extracted lists of marker genes from multiple sources for users to search. While having a consolidated source of marker gene lists is helpful, researchers must still sort through data to identify which marker genes are robust and relevant to their analyses. For example, identifying species-specific markers, markers consistent across samples, and markers able to be detected in 3′-sequencing methods, would require the users to manually identify marker genes fitting their data criteria. Therefore, the primary focus of this manuscript is to provide a resource to document the marker genes that were consistently identified across species, samples, sequencing technologies, and sources.

To identify consistent marker genes for specific cell types, we manually curated results from publications that performed large-scale statistical analyses on pure cell populations via scRNA-seq or Fluorescence-activated cell sorting (FACS) methodologies. We focused on publications using expression data from mice and/or human untreated, non-disease samples. Next, the extracted gene information was standardized to known ontologies, cellular hierarchy information was incorporated, and a marker gene score algorithm to identify marker genes consistent across sources, samples, and species was developed. Two MySQL databases were generated to store: 1) the standardized, manually curated statistical results and metadata and 2) the robust marker genes, while an R Shiny reactive user-interface is provided to access the data. The development of the publicly accessible GeneMarkeR database and user-interface is described in this manuscript.

MATERIALS AND METHODS

Data Extraction

Data curated for the database focused on publications concentrated on performing statistical analyses to identify cell type-specific marker genes in their samples. There were 25 unique marker gene analyses from these publications that either: 1) used scRNA-seq expression data, 2) used RNA-seq or microarray expression data collected from pure cell populations, or 3) came from collaborators sharing highly validated (i.e., prototypical) marker genes. Additional publications were evaluated; however, these were filtered out as the exclusive focus was on naïve (i.e., non-treated, non-disease) mouse and human samples. The marker genes, cell types and full statistical results were manually extracted from figures, supplemental data, and text of publications, or directly from the author to ensure data integrity. In a few cases only the significant marker gene results were available from the author, not the full statistical output. Additional contextual data (i.e., sample type, species, gene expression method, statistical method, relevant statistical cutoffs) were collected from each source. For publications that used scRNA-seq data, prototypical marker genes, marker genes the authors used to annotate their cell types for each cell population, were extracted. These prototypical marker genes are generally highly validated, well-accepted genes used to annotate cell types prior to performing novel marker gene identification.

Ontology Standardization

To enable mouse-human comparison across the same genomic entity (i.e., genes, miRs, lncRNAs), Mouse Genome Informatics (MGI) and Entrez mouse-human ortholog information were used to map genomic entity information. Genomic entities for mouse (assembly GRCm39) and human (assembly GRCh38.p13) were standardized using gene symbols and unique identifiers from both Entrez and Ensembl. A unique key (GeneID) was generated to identify each unique mouse-human ortholog pair, or when no ortholog is described, to denote the mouse or human-specific genomic entity. A total of 21,012 unique genomic entities were included in the analysis. Genomic entities are referred to as genes in the Figures and Tables for readability as genes comprise most of the genomic entities.

The 120 distinct cell types extracted from the publications were mapped to Cell Ontology terms using EMBL-EBI's Ontology Lookup Service and Ontobee. Additional cell types were added to the network structure to ensure specific cell types accurately mapped back to parent nodes (i.e., naïve cell and somatic cell). Redundant terms (i.e., cell types that mapped in multiple branches) were pruned by removing cyclic relationships manually. Intermediate nodes that lacked branching and did not add value to the classification were manually removed. Intermediate nodes with branches were retained as these are crucial to build out the tree as cell types from new datasets are added. The cell type hierarchy of Cell Ontology was built via the JavaScript package "visNetwork" implemented in R with an abbreviated version shown in **Figure 1A**. The cell hierarchy enables us to consider if genes were specific for higher-level cell type terms vs. cell subtypes.

Marker Gene Score

To compare disparate marker gene statistics across publications, each statistical endpoint from a source was normalized between 0 and 1. The midpoint (i.e., 0.5) was set as the author provided statistical significance cut-off. For example, in **Supplementary Figure S1** the example Source 1 had two distinct statistical endpoints: 1) log fold change enrichment score, and 2) adjusted *p*-value. The log fold change enrichment score ranges from −9 to 0 where the more negative the result, the more significant. For log fold change enrichment score, these authors considered results less than or equal to −2 to be statistically significant; therefore, −2 is set at a marker gene score of 0.5 while values between −9 and −2 are scaled between 1 and 0.5, respectively and values between −2 and 0 are scaled between 0.5 and 0, respectively. The adjusted *p*-value for Source 1 ranged from 0 to 1, with increasing significance closer to 0. For adjusted *p*-value, these authors considered results less than or equal to 0.05 to be statistically significant; therefore, 0.05 is set at a marker gene score of 0.5 while values between 0 and 0.05 are scaled between 1 and 0.5, respectively and values between 0.05 and 1 are scaled between 0.5 and 0, respectively. The preliminary scores were averaged across the source per gene-cell type pair to calculate a marker gene score for each unique

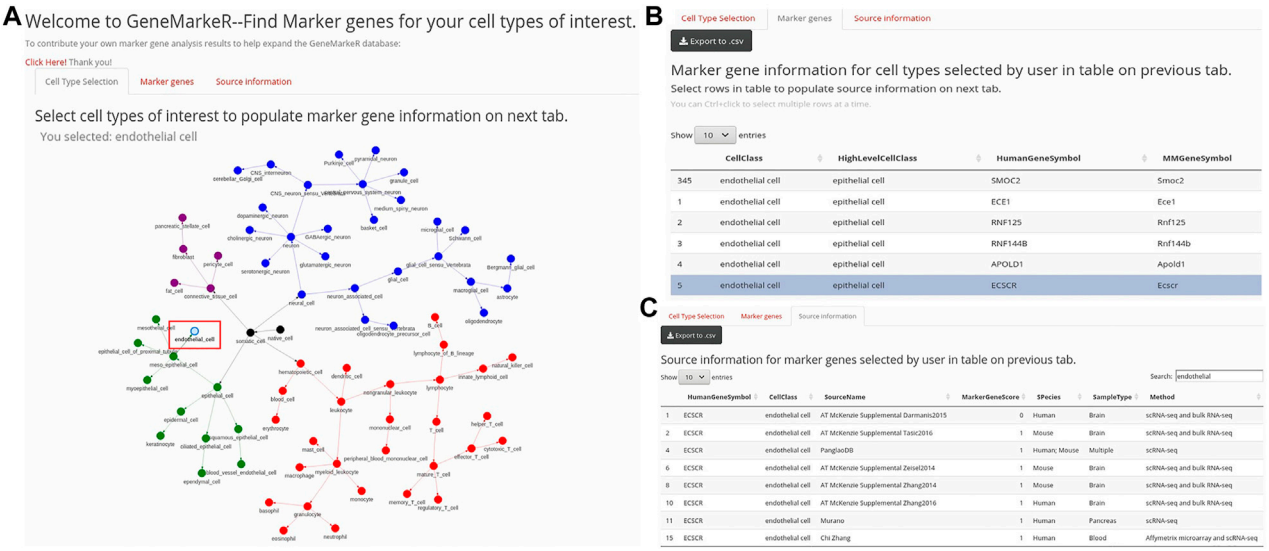


FIGURE 1 | GeneMarkeR user interface. **(A)** "Cell Type Selection" tab enables selection of cell types from a drop-down menu (not shown) and a hierarchical cell network (abbreviated network shown here). Upon selection of cell types, the "Marker genes" tab is populated. **(B)** "Marker genes" tab displays marker genes and species specificity information from GeneMarker.db filtered to the user-selected cell types. Upon selection of marker genes of interest, the "Source information" tab is populated. **(C)** "Source information" tab displays source, sample and statistical data from CellSearcheR.db filtered to the user-selected marker genes. All GeneMarkeR data can be exported to CSV.

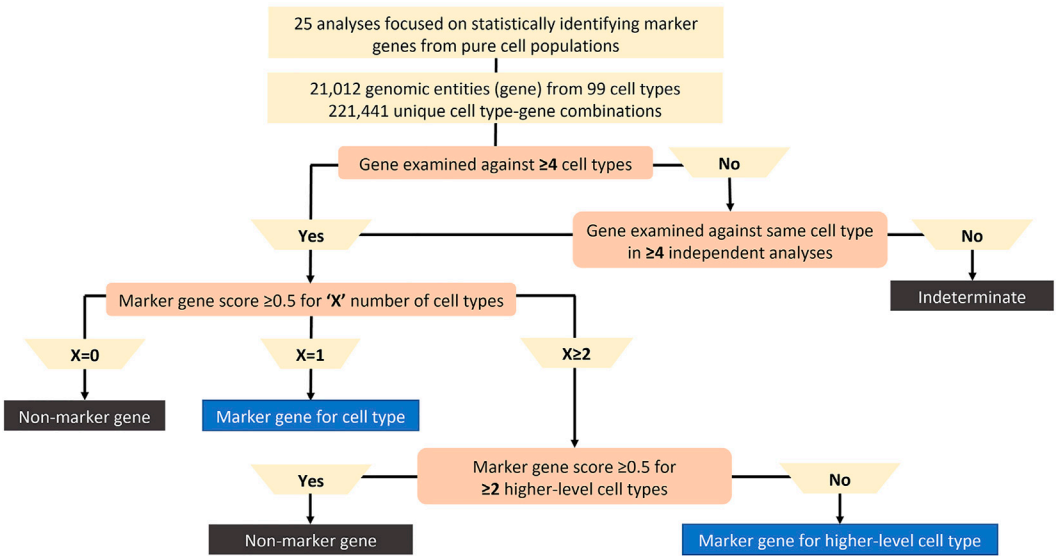


FIGURE 2 | Marker gene classification algorithm. The algorithm classifies genes with <4 cell types as Indeterminate due to insufficient data to analyze the gene unless the gene was examined against the same cell type in ≥4 separate analyses. Significant marker gene score (i.e., at least 0.5) for at least two-thirds of publications considered for each gene-cell type pair. The number of significant gene-cell type pairs per a given gene is "X". 1) If X = 0, the gene is not a marker gene for any cell type, 2) If X = 1, the gene is a marker gene for one cell type, 3) If X ≥ 2, then the gene is a marker gene for more than one cell type. When X ≥ 2, if the gene is marker gene for multiple cell types from the same higher-level cell type (ex: connective tissue cells), then the gene is a marker gene for the higher-level cell type. When X ≥ 2, if the gene is marker gene for multiple cell types from at least 2 higher-level cell types (ex: connective tissue cells and T cells), then the gene is a non-marker gene due to lack of specificity across publications.

gene-cell type-source combination as is shown in **Supplementary Figure S1**. For example, in **Supplementary Figure S1**, if a unique gene-cell type pair are reported to have a log fold change enrichment score of −9 and an adjusted *p*-value of 0.05, then the preliminary scores of 1 and 0.5, respectively, would be averaged, resulting in a marker gene score of 0.75 for that

gene-cell type pair in Source 1. A marker gene score of 1 indicates strong evidence that a gene is a marker gene for a given cell type from that source, while 0 indicates little to no evidence for supporting this relationship.

Marker Gene Score Algorithm

To classify whether a gene was specific across samples, species and sources for a given cell type, a simple marker gene classification algorithm was developed as shown in **Figure 2**. Genes reported in fewer than 4 cell types were labelled as Indeterminate due to insufficient data to determine specificity across multiple cell types. As highly specific genes may not be expressed in other cell types accounting for reporting in fewer than 4 cell types, a subset of genes categorized as Indeterminate had to be reclassified. Therefore, genes originally classified as Indeterminate that were analysed in the same cell type across at least 4 separate sources for common cell types or 2 separate sources for rare cell types (e.g., pancreatic epsilon cell) were reclassified and were included in the next classification steps.

Next, the number of cell types (X) with an average marker gene score of ≥0.5 across publications were counted for each gene. If X = 0 for an individual gene, that gene was not considered a marker gene for any cell type (i.e., a non-marker gene). If X = 1 for an individual gene, that gene was significant for a single cell type across sources, so it was classified as a marker gene for that cell type. To ensure genes were specific for a limited number of cell types, each gene was restricted to be considered a marker gene for a maximum of 2 cell types. To ensure this cut-off was achieved, if $X \geq 2$ for an individual gene, the number of higher-level cell types (Y) were considered. If $Y < 2$ for an individual gene, then the gene was a marker gene for the higher-level cell type. If $Y \geq 2$ for an individual gene, the gene would be considered in most cases as a non-marker gene since it was not specific across publications. As each gene was restricted to a maximum of 2 cell types for which it was specific, genes exceeding this are labelled non-marker genes.

Therefore, using our algorithm cut-off X, we first check if the gene is specific for the more granular cell subtypes. If $X < 2$, then the gene is subtype specific, thus specific for that cell subtype and for any higher-level cell types in the hierarchical tree branch. While we count this gene as specific for 1 cell type subtype, the specificity relationship is propagated up the branch meaning the gene is also specific for higher level cell types in that branch. If $X \geq 2$, then we check if those cell subtypes fall under the same higher-level cell type by looking at the built hierarchical ontology tree structure. This is where the higher-level threshold of Y comes into play. If a gene is found to be specific for multiple cell subtypes (i.e., $X \geq 2$) and those cell subtypes belong to the same higher level cell subtype, then the gene is a marker gene for the higher-level cell type, but NOT for the subtypes. Species specificity for a marker gene required a 3-fold difference in median marker gene score between species with the median exceeding 0.5 for at least one of the species.

Database Design and Web Interface

There are two databases behind GeneMarkeR shown in **Figure 3**, they are both implemented in MySQL to ensure data integrity, standardization, and ease of data updates over time. CellSearcheR.db consists of over 261,000 rows of data extracted across 15 publications and 2 datasets from collaborators comprising a total of 25 unique marker gene analyses. CellSearcheR.db was processed through the algorithm described in *Materials and Methods Marker gene score* to create GeneMarkeR.db, which stores gene-cell type relationships for the algorithm identified marker genes.

An R Shiny tool hosted on the IU Precision Health Initiative server enables access and extraction of both CellSearcheR.db and GeneMarkeR.db databases. As is shown in **Figures 1A–C**, the R Shiny tool has reactive programming built-in, so when the user selects cell types, this accesses GeneMarkeR.db to populate the marker gene tab with algorithm-derived marker genes for their cell types of interest. User selection of genes of interest on the marker gene tab reactively retrieves the standardized, raw CellSearcheR.db marker gene score and statistical data for each of those genes.

A link (https://redcap.uits.iu.edu/surveys/?s=XEAFCX4LC7) is provided on the web interface to a user submission form where researchers can submit their marker gene analysis data. The online form provides the results in a standardized CSV output to enable easy standardization and addition to CellSearcheR.db. In addition, marker gene analyses from new publications can also be manually extracted and standardized to update CellSearcheR.db with new data. The marker gene score algorithm is then used to process all the data in CellSearcheR.db to update the results in GeneMarkeR.db. Therefore, the process ensures updatability of the databases and web interface over time from user submission and manual extraction from new publications.

RESULTS

In total, 25 unique marker gene analyses of 9 distinct specimen types (blood, bone marrow, brain, heart, kidney, lung, pancreas, and tonsil) and additional cross-specimen sample types were identified that met the criteria specified in the *Materials and Method* section. The 261,000 rows of standardized marker gene statistical data extracted from the 25 analyses were stored in the CellSearcheR.db. As is shown in **Figure 3**, the CellSeatcheR.db data are analyzed in the marker gene classification algorithm detailed in **Figure 2** to identify the marker genes that are then stored in GeneMarkeR.db. The information housed in each database is shown in **Figure 3** and the data from both MySQL databases are used to generate the GeneMarkeR Tool R Shiny interface.

The 3,936 genomic entities that could not be automatically or manually mapped to a current gene annotation were excluded, leaving 21,012 genomic entities for the analysis. There were over 120 distinct cell types (including higher level cell types) with 221,441 unique gene-cell type combinations considered in the marker gene analysis. The final analysis of standardized marker gene results identified 2,464 genes as specific for one or two cell types with 2,746 total marker gene pairs as 281 genes were specific for two cell types. 7,283 genes were classified as non-marker genes, 10,465 were classified as indeterminate due to sparse data and the remainder were a mix of non-marker gene and indeterminate. The number of genes identified as a marker gene

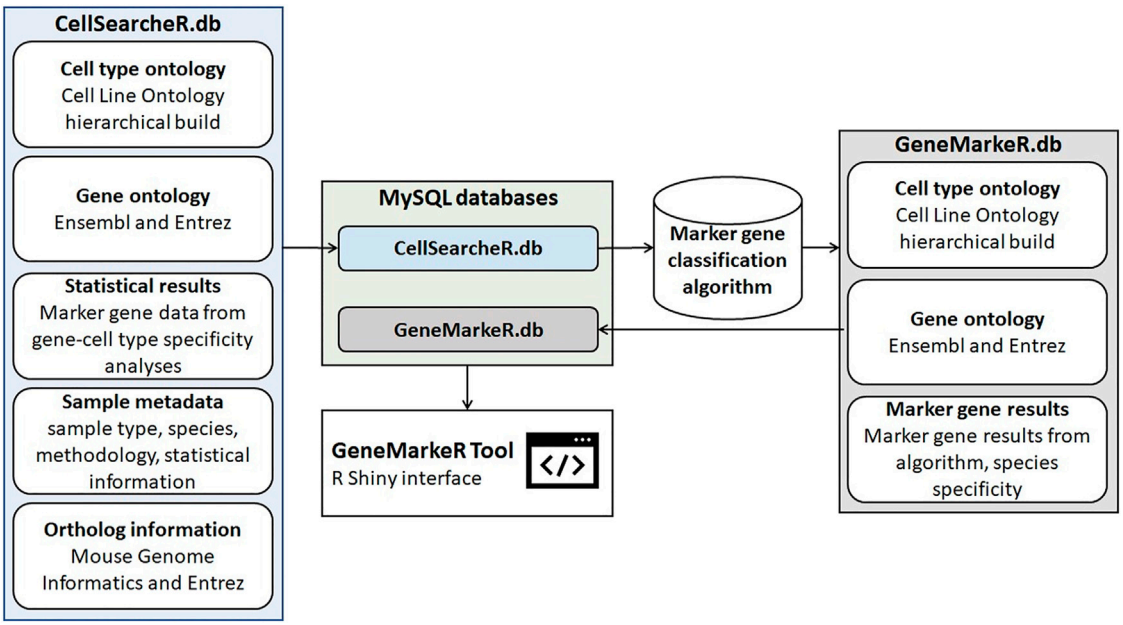


FIGURE 3 | Schema for generation of GeneMarkeR. The CellSearcheR database (CellSearcheR.db) integrates standardized cell type, gene and statistical data from publications performing marker gene identification analyses on non-treated, non-disease mouse and human samples. Data was manually extracted, cell types mapped to the Cell Line Ontology hierarchical ontology, genes mapped to Ensembl and Entrez identifiers and statistical data normalized to a marker gene score. CellSearcheR.db is processed by the marker gene classification algorithm detailed in **Figure 2**. The GeneMarkeR database (GeneMarkeR.db) integrates the standardized cell type, gene and marker gene results from the algorithm output. CellSearcheR.db and GeneMarkeR.db are MySQL databases that are accessed via an R Shiny user interface called GeneMarkeR.

analyzed at that cell type (dark blue) out of all genes analyzed at that cell type (length of bar) is shown in **Figure 4A**. There were 68 cell types with marker genes identified out of the 120 cell types extracted from the 25 unique marker gene analyses. Out of the 2,746 marker genes, 80% of those were classified as a specific cell type and 20% were classified as a higher-level cell type. Filtering to the marker genes from **Figure 4A** (dark blue) we get **Figure 4B** where marker genes are categorized based on whether the gene is specific for that cell type (light blue) or a higher-level cell type (purple). For example, there were 5,000 genes analyzed against fibroblasts with approximately 500 being identified as marker genes for fibroblasts and 400 being identified as marker genes for a higher-level cell type (i.e., connective tissue cell).

In CellMarker there are an average of 2.2 sources supporting marker genes in normal tissue samples with 55% of marker genes supported by a single source. In the GeneMarkeR.db database there are 4.5 sources on average supporting a gene being a marker gene for a certain cell type in our database with only 4 (0.1%) marker genes supported by a single source. These 4 cases were due to the gene being a higher-level marker gene in the cell ontology and the individual publications having at least 4 distinct cell types to support that re-classification.

DISCUSSION

The analysis described here focused on mouse and human as these two species comprise most marker gene data analyses. Non-treated and non-disease samples were evaluated to study the naïve state of cell identity. This enables future analyses to delve into the impact that disease and treatment may have on cell identity markers. After extracting data from public datasets meeting these criteria, data standardization was addressed. Due to differences in genome annotations, sources of gene symbols, and naming conventions across publications, not all genes could be automatically mapped. Therefore, 15% of the gene symbols were manually mapped to current genome assembly GRCm39 for mouse and GRCh38.p13 for human. Genes that existed in earlier genome annotations but have since been discontinued in current mouse and human reference genomes were removed from the analysis.

The ontology standardization of cell type started with mapping cell types from the publications to Cell Ontology. Nodes of these cell types and their higher-level cell types were connected by building the network backwards from the most specific cell types up to the highest-level parent nodes (i.e., naïve cell or somatic cell). As is described in *Materials and Methods Ontology standardization*, the hierarchy was manually pruned to remove redundancy and circular relationships, while maintaining intermediate cell type nodes to ensure new cell types could be connected in the future. In a handful of cases nodes were manually adjusted to ensure biological relevance and consistency. The higher-level cell types were then added to the database to improve the marker gene score algorithm.

Due to differences in statistical methods, endpoints and significance cut-offs, the marker gene score was calculated to enable normalization and comparison across publications. Using the median and average marker gene score we used the marker gene score algorithm to identify marker gene, higher-level marker

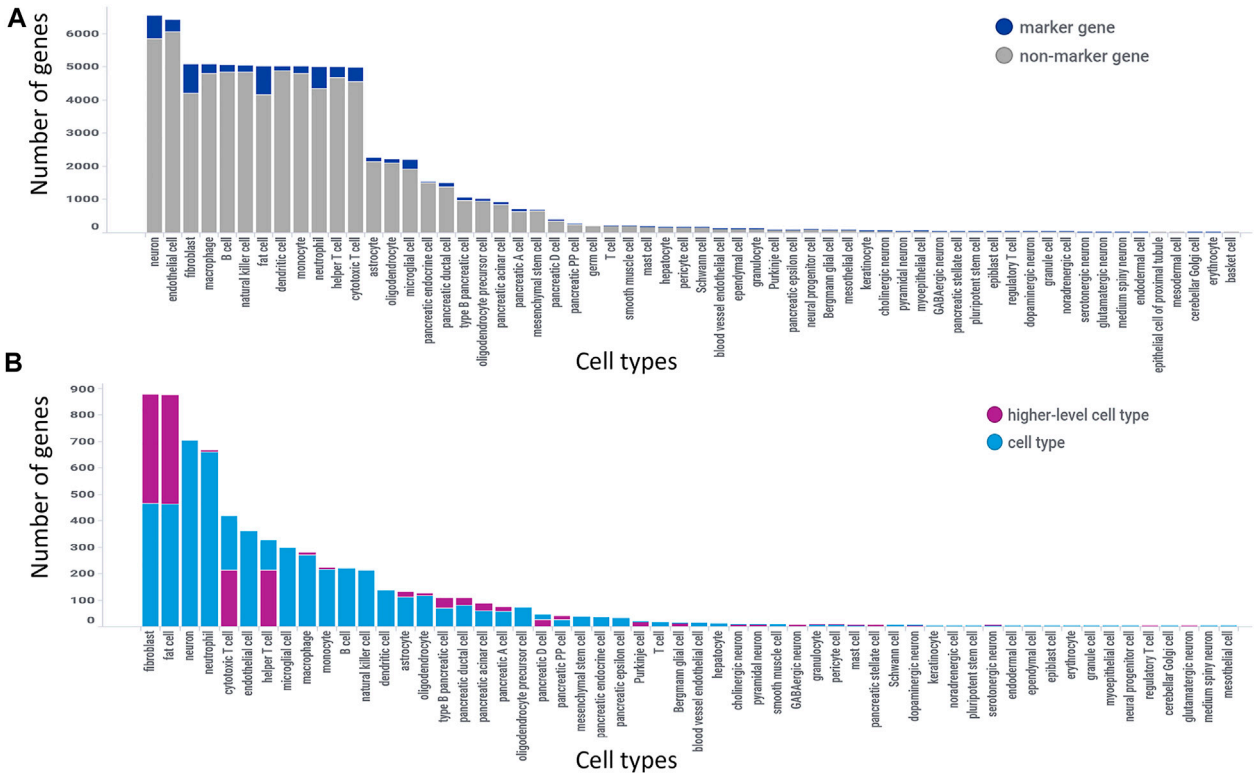


FIGURE 4 | GeneMarkeR summary statistics. **(A)** Number of marker genes identified (dark blue) out of all genes analyzed per cell type (length of bar). **(B)** Number of marker genes reclassified as a higher-level cell type marker (purple) and number of marker genes classified as cell type-specific marker (light blue) out of all marker genes identified for that cell type (length of bar). **Figure 2B** is filtered to the marker genes, i.e., dark blue bars in **figure 2A**. Both figures are sorted from largest to smallest number of genes.

gene, non-marker gene and indeterminate calls for each cell type-gene combination across sources. While approximately 80% of gene-cell type pairs could be automatically annotated by following the algorithm, genes with more than 2 higher-level cell types had to be manually checked to determine if those higher-level cell types were from the same branch of the hierarchical cell map or from a previously pruned branch. For example, microglial cell can be connected to multiple branches (i.e., glial cell, macrophage, and myeloid cell, etc. . .), so the manual mapping would reconsider these additional connections and higher-level cell types in context of all data for that gene-cell type pair.

While CellMarker is a great source of marker gene annotations from normal and disease samples, the database described in this manuscript provides an improvement in marker gene identification for normal mouse and human samples. An advantage of this algorithm over previously published analyses is the greater amount of data supporting each marker gene call. Identifying genes that are considered as marker genes across multiple sources in CellMarker requires users to perform their own analysis of the data, whereas GeneMarkeR provides the user with that information. In addition, unlike CellMarker, GeneMarkeR considers the difference and overlap between mouse and human enabling species-specific gene markers to be included or excluded. Finally, due to the inclusion of hierarchical cell ontology in GeneMarkeR, 538 genes were more accurately reclassified as being specific for a higher-level cell type rather than the original publication cell type, which is not considered in CellMarker.

In conclusion, data were first manually extracted from publicly available marker gene analyses and hierarchical ontology standardization was applied to create CellSearcheR.db. Next, GeneMarkeR.db was developed using a novel algorithm that considers marker gene score to identify marker genes specific across species, samples, and methodology. Finally, an R Shiny user interface was developed (GeneMarkeR) that pulls from CellSearcheR.db and GeneMarkeR.db using reactive programming. The GeneMarkeR tool provides highly validated, consistent marker genes and species specificity information to enable improved scRNA-seq cell type identification over existing databases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author. GeneMarkeR is freely available at https://shiny.ph.iu.edu/GeneMarkeR/ access on the web with all major browsers supported.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We thank IU Precision Health Initiative facility for hosting the R Shiny application as well as Peter Barker for setting up the MySQL server.

REFERENCES

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., et al. (2019). Reference-based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. *Nat. Immunol.* 20 (2), 163–172. doi:10.1038/s41590-018-0276-y

Franzén, O., Gan, L.-M., and Björkegren, J. L. M. (2019). PanglaoDB: a Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data. *Database* 2019, baz046. doi:10.1093/database/baz046

Mancarci, B. O., Toker, L., Tripathy, S. J., Li, B., Rocco, B., Sibille, E., et al. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eNeuro* 4 (6), 0212–0217. doi:10.1523/ENEURO.0212-17.2017

Saviano, A., Henderson, N. C., and Baumert, T. F. (2020). Single-cell Genomics and Spatial Transcriptomics: Discovery of Novel Cell States and Cellular Interactions in Liver Physiology and Disease Biology. *J. Hepatol.* 73, 1219–1230. doi:10.1016/j.jhep.2020.06.004

Skelly, D. A., Squiers, G. T., McLellan, M. A., Bolisetty, M. T., Robson, P., Rosenthal, N. A., et al. (2018). Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart. *Cel Rep.* 22 (3), 600–610. doi:10.1016/j.celrep.2017.12.072

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a Manually Curated Resource of Cell Markers in Human and Mouse. *Nucleic Acids Res.* 47 (D1), D721–D728. doi:10.1093/nar/gky900

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763431/full#supplementary-material

Supplementary Figure 1 | Marker gene score calculation for an individual source/publication. For each marker gene statistic in an individual source, the statistical results are scaled from 0 (low/no evidence marker gene) to 1 (high evidence marker gene) using the author specified statistical significance cutoffs. All gene-cell type pairs in that source receive a preliminary score for each endpoint. Those preliminary scores are then averaged to calculate the marker gene score for a gene-cell type pair in an individual source. A given gene-cell type pair will have a different marker gene score for each publication to enable comparison of gene specificity for that cell type across sources.

Conflict of Interest: Author BP was employed by company Eli Lilly and Company, United States.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JL declared a past co-authorship with one of the authors YL to the handling editor.

frontiers in Genetics

METHODS
published: 07 December 2021
doi: 10.3389/fgene.2021.766405

tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data

Johannes Ostner[1,2], *Salomé Carcy*[2,3†] *and Christian L. Müller*[1,2,4*]

[1]*Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany,* [2]*Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany,* [3]*Department of Biology, École Normale Supérieure, PSL University, Paris, France,* [4]*Center for Computational Mathematics, Flatiron Institute, New York, NY, United States*

Accurate generative statistical modeling of count data is of critical relevance for the analysis of biological datasets from high-throughput sequencing technologies. Important instances include the modeling of microbiome compositions from amplicon sequencing surveys and the analysis of cell type compositions derived from single-cell RNA sequencing. Microbial and cell type abundance data share remarkably similar statistical features, including their inherent compositionality and a natural hierarchical ordering of the individual components from taxonomic or cell lineage tree information, respectively. To this end, we introduce a Bayesian model for **t**ree-aggregated **a**mplicon and **s**ingle-**c**ell **co**mpositional **d**ata **a**nalysis (tascCODA) that seamlessly integrates hierarchical information and experimental covariate data into the generative modeling of compositional count data. By combining latent parameters based on the tree structure with spike-and-slab Lasso penalization, tascCODA can determine covariate effects across different levels of the population hierarchy in a data-driven parsimonious way. In the context of differential abundance testing, we validate tascCODA's excellent performance on a comprehensive set of synthetic benchmark scenarios. Our analyses on human single-cell RNA-seq data from ulcerative colitis patients and amplicon data from patients with irritable bowel syndrome, respectively, identified aggregated cell type and taxon compositional changes that were more predictive and parsimonious than those proposed by other schemes. We posit that tascCODA[1] constitutes a valuable addition to the growing statistical toolbox for generative modeling and analysis of compositional changes in microbial or cell population data.

Keywords: bayesian modeling, dirichlet multinomial, microbiome data, single-cell data, spike-and-slab lasso, tree aggregation, differential abundance testing

OPEN ACCESS

Edited by:
Himel Mallick,
Merck, United States

Reviewed by:
Boyu Ren,
Dana–Farber Cancer Institute, United States
Thomas P. Quinn,
Deakin University, Australia
Siyuan Ma,
University of Pennsylvania, United States

***Correspondence:**
Christian L. Müller
christian.mueller@helmholtz-muenchen.de

***Present Address:**
Salomé Carcy,
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY, United States

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 August 2021*
Accepted: *01 November 2021*
Published: *07 December 2021*

Citation:
Ostner J, Carcy S and Müller CL (2021) tascCODA: Bayesian Tree-Aggregated Analysis of Compositional Amplicon and Single-Cell Data. Front. Genet. 12:766405. doi: 10.3389/fgene.2021.766405

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have fundamentally transformed our ability to quantitatively measure the molecular make-up of single cells (Shalek et al., 2013), tissues (Regev et al., 2017; Karlsson et al., 2021), organs (He et al., 2020), as well as microbiome compositions in and on the human body (Human Microbiome Project Consortium, 2012). Single-cell RNA

[1]Available at https://github.com/bio-datascience/tascCODA.

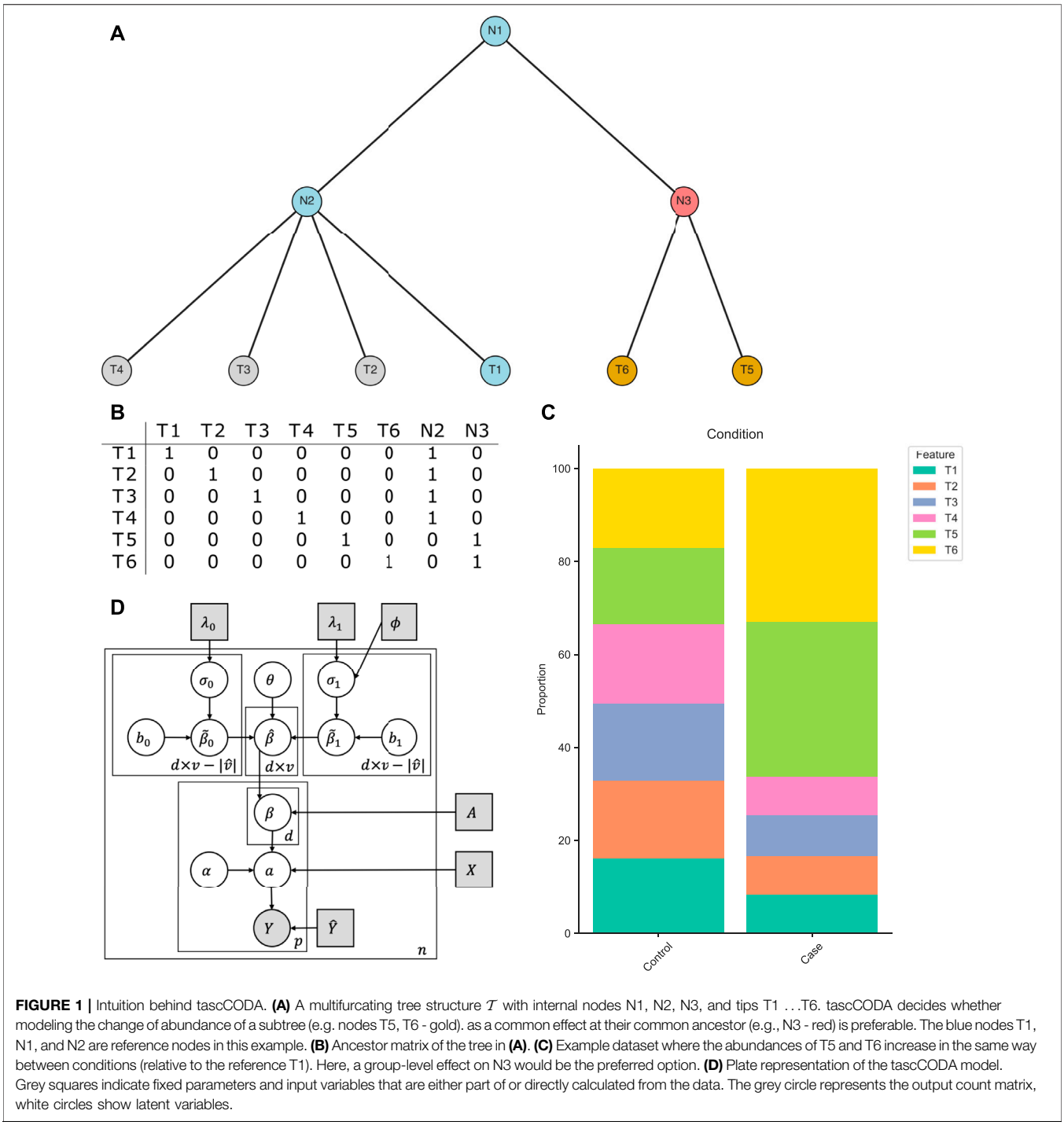
sequencing (scRNA-seq) (Tang et al., 2009; Shalek et al., 2013; Macosko et al., 2015) has become the key technology for recording the transcriptional profiles of individual cells across different tissue types (Regev et al., 2017) and developmental stages (Griffiths et al., 2018), and for determining cell type states and overall cell type compositions (Trapnell, 2015). Cell type compositions provide informative and interpretable representations of the noisy high-dimensional scRNA-seq data and are typically derived from clustering characteristic gene expression patterns in each cell (Duò et al., 2018; Traag et al., 2019), followed by analysis of the expression levels of marker genes (Luecken and Theis, 2019). As a by-product, these workflows also yield a hierarchical grouping of the cell types, either derived from the clustering procedure or determined by known cell lineage hierarchies. Determining changes in cell type populations across conditions can give valuable insight into the effects of drug treatment (Tsoucas et al., 2019) and disease status (Smillie et al., 2019), among others.

Complementary to scRNA-seq data collection, amplicon or marker-gene sequencing techniques provide abundance information of microbes across human body sites (Human Microbiome Project Consortium, 2012; Lloyd-Price et al., 2017; McDonald et al., 2018). Current estimates suggest that the human microbiome, i.e., the collection of microbes in and on the human body, outnumber an individual's somatic and germ cells by a factor of 1.3–10 (Turnbaugh et al., 2007; Sender et al., 2016). Starting from the raw read counts, amplicon data are typically summarized in count abundance tables of operational taxonomic units (OTUs) at a fixed sequence similarity level or, alternatively, of denoised amplicon sequence variants (ASVs). The marker genes also allow taxonomic classification and phylogenetic tree estimation, thus inducing a hierarchical grouping of the taxa. To reduce the dimensionality of the data set and guard against noisy and low count measurements, the taxonomic grouping information is often used to aggregate the data at a fixed taxonomic rank, e.g., the genus or family rank. Shifts in the population structure of taxa have been implicated in the host's health and have been associated with various diseases and symptoms, including immune-mediated diseases (Round and Palm, 2018), Crohn's disease (Gevers et al., 2014), and Irritable Bowel Syndrome (IBS) (Ford et al., 2017).

In the present work, we exploit the remarkable similarities between scRNA-seq-derived cell type data and amplicon-based microbial count data and propose a statistical generative model that is applicable to both data modalities: the Bayesian model for **t**ree-aggregated **a**mplicon and **s**ingle-**c**ell **CO**mpositional **D**ata **A**nalysis, in short, tascCODA. Our model assumes that count data are available in the form of a $n \times p$-dimensional count matrix Y containing the counts of p different cell types or microbial taxa in n samples, a covariate matrix $n \times d$-dimensional X carrying metadata or covariate information for each sample, and a tree structure with p leaves that imposes a hierarchical order on the count data Y. Since both amplicon and scRNA-seq technologies are limited in the amount of material that can be processed in one sample, the total number of counts in rows of Y do not reflect total abundance measurements of the features but rather relate to the efficiency of the sequencing experiment itself (Gloor et al., 2017). This implies that the counts only carry relative abundance information, making them essentially compositional data (Aitchison, 1982).

tascCODA is a fully Bayesian model for tree-aggregated modeling of count data and is a natural extension of the scCODA model, recently introduced for compositional scRNA-seq data analysis (Büttner et al., 2020). At its core, tascCODA models the count data Y via a Dirichlet Multinomial distribution and associates count data and covariate information via a log-link function. To encourage sparsity in the underlying associations between the covariates and the hierarchically grouped features, tascCODA exploits recent ideas from tree-guided regularization and the spike-and-slab LASSO (Ročková and George (2018)). This allows tascCODA to perform tree-guided sparse regression on compositional responses with any type or number of covariates. In particular, in the presence of a single binary covariate, e.g., a condition indicator, tascCODA allows to perform Bayesian differential abundance testing. More generally, however, tascCODA enables to determine how host phenotype, such as disease status, host covariates such as age, gender, or an individual's demographics, or environmental factors jointly influence the compositional counts. Finally, incorporating tree information into the inference allows tascCODA to not only identify associations between individual features, but also entire groups of features that form a subset of the tree.

tascCODA complements several recent statistical approaches, in particular, from the field of microbiome data analysis, some of which also use the concept of tree-guided models. Chen and Li (2013) were among the first to use the sparse Dirichlet-Multinomial model to connect compositional count data with covariate information in a penalized maximum-likelihood setting. Wadsworth et al. (2017) were the first to use a similar model in a Bayesian setting. Both adaANCOM (Zhou C. et al. (2021)) and the Logstic-tree normal model (Wang et al. (2021)) use the Dirichlet-tree (multinomial) model (Wang and Zhao (2017)) to determine differential abundance of microbial taxa via a product of Dirichlet distributions at each split. The PhILR model (Silverman et al., 2017) uses the phylogenetic tree of a microbial community to compute an isometric logratio transform with interpretable balances. Furthermore, there are recent advances in constructing optimal hierarchical partitions of HTS data and to predict variables of interest from them (Quinn and Erb, 2019; Gordon-Rodriguez et al., 2021), that do not rely on pre-defined trees, but rather structure the data in the best way to be predictive of the outcome. These methods restrict themselves, however, to fully binary trees. On the other hand, the trac method (Bien et al., 2021) uses tree-guided regularization (Yan and Bien, 2021) in a maximum-likelihood-type framework to predict continuous outcomes from compositional microbiome data.



	T1	T2	T3	T4	T5	T6	N2	N3
T1	1	0	0	0	0	0	1	0
T2	0	1	0	0	0	0	1	0
T3	0	0	1	0	0	0	1	0
T4	0	0	0	1	0	0	1	0
T5	0	0	0	0	1	0	0	1
T6	0	0	0	0	0	1	0	1

FIGURE 1 | Intuition behind tascCODA. **(A)** A multifurcating tree structure $\mathcal{T}$ with internal nodes N1, N2, N3, and tips T1 . . .T6. tascCODA decides whether modeling the change of abundance of a subtree (e.g. nodes T5, T6 - gold). as a common effect at their common ancestor (e.g., N3 - red) is preferable. The blue nodes T1, N1, and N2 are reference nodes in this example. **(B)** Ancestor matrix of the tree in **(A)**. **(C)** Example dataset where the abundances of T5 and T6 increase in the same way between conditions (relative to the reference T1). Here, a group-level effect on N3 would be the preferred option. **(D)** Plate representation of the tascCODA model. Grey squares indicate fixed parameters and input variables that are either part of or directly calculated from the data. The grey circle represents the output count matrix, white circles show latent variables.

In its present form, the Bayesian model behind tascCODA is ideally suited for data sets of moderate dimensionality, typically $p < 100$, yet can handle extremely small sample sizes n. Since amplicon datasets are usually high-dimensional in the number of taxa and exhibit high overdispersion and excess number of zeros, we focus on the analysis of genus-level microbiome data. In the context of cell type compositional data, on the other hand, often only very few replicate samples are available (Büttner et al., 2020). Here, tascCODA can leverage well-calibrated prior information to operate in low-sample regimes where frequentist methods likely fail.

The remainder of the paper is structured as follows. In the next section, we introduce the tascCODA model and describe the computational implementation. In **Section 3**, we describe and discuss synthetic data benchmarks and provide two real-world applications, on human single-cell RNA-seq data from ulcerative

colitis patients and amplicon data from patients with irritable bowel syndrome. Finally, we summarize the key points in **Section 4** and present considerations about future extensions of the method. A flexible and user-friendly implementation of tascCODA is available in the Python package *tascCODA*[2]. All results in this paper are fully reproducible and available on Zenodo[3].

2 MATERIALS AND METHODS

2.1 Model Description

We start with formally describing the problem at hand. Let $Y \in \mathbb{R}^{n \times p}$ be a count matrix describing n samples from p features (e.g., cell types, microbial taxa, etc.), and $X \in \mathbb{R}^{n \times d}$ be a matrix that contains the values of d covariates of interest for each sample. Due to the technical limitations of the sampling procedure, the sum of counts in each sample, $\bar{Y}_i = \sum_{j=1}^{p} Y_{i,j}$ must be seen as a scaling factor, making the data compositional (Gloor et al. (2017)). Additionally, the features described by Y are hierarchically ordered by a tree $\mathcal{T}$ with p leaves and t internal nodes, resulting in a total number of $v = p + t$ nodes in $\mathcal{T}$ (**Figure 1A**). Such tree structures are usually motivated by taxonomy (McDonald et al., 2012; Quast et al., 2013), determined by phylogenetic similarities (Schliep, 2010), or obtained via serial binary partitions (Quinn and Erb, 2019). The tree can further be bifurcating or multifurcating, thus internal nodes may have two or more descendants.

$\mathcal{T}$ can be fully characterized by a binary ancestor matrix $A \in \{0,1\}^{p \times v}$. Hereby, each row of A stands for a feature or leaf node of $\mathcal{T}$, the first p columns also denote the leaves of the tree, and the last t columns represent the internal nodes. The entries $A_{j,k}$ are 1, if column k corresponds either to feature j ($j = k$) or to one of its parents, otherwise it is 0 (**Figure 1B**):

$$A_{j,k} = \begin{cases} 1 & \text{if } j = k \text{ or } k \text{ is ancestor of } j \\ 0 & \text{else.} \end{cases}$$

Our goal is to determine how changes in abundance of features (leaves of $\mathcal{T}$) are associated with the covariates in X, and select a sparse set of the most important covariate-feature effects. To achieve an even more parsimonious result, we further determine whether groups of features that form subtrees of $\mathcal{T}$ are affected by the conditions in the same manner (**Figure 1A**), and model them with a common effect if possible. This group-wise modeling step not only gives an accurate, yet easy to interpret description of the changes in the feature composition, but can also reveal shared traits among structural subgroups of features that might be missed in analyses that do not take the tree structure into account.

2.1.1 Core Model With Tree Aggregation

tascCODA posits a Dirichlet-Multinomial model for $Y_{i,\cdot}$ for each sample $i \in 1\ldots, n$, thus accounting for the compositional nature of the count data. The covariates are associated with the features through a log-linear relationship. We put uninformative Normal priors on the base composition α, which describes the data in the case $X_{i,\cdot} = 0$:

$$Y_i \sim \text{DirMult}\left(\bar{Y}_i, \mathbf{a}(X)_i\right) \tag{1}$$

$$\log\left(\mathbf{a}(X)\right)_i = \alpha + X_{i,\cdot}\beta \tag{2}$$

$$\alpha_j \sim \mathcal{N}(0, 10) \qquad \forall j \in [p]. \tag{3}$$

The total count $\bar{Y}_i$ is directly inferred from the data for each sample. The effect of the lth covariate on the jth feature is therefore given by $\beta_{l,j}$.

We now use a variant of the tree-based penalty formulation of Yan and Bien (2021) to model common effects at each internal node of $\mathcal{T}$ in addition to the effects on the leaves. We define a node effect matrix $\hat{\beta} \in \mathbb{R}^{d \times v}$ and associate aggregations on internal nodes with the correct tips by multiplying with the ancestor matrix A:

$$\beta = \hat{\beta} A^T \tag{4}$$

To illustrate the intuition behind this step, we consider an example based on the tree in **Figure 1A**. In a binary covariate setting, the features T1-T6 are uniformly distributed in the control population, while in the case population, the abundance of features T5 and T6 (with respect to feature T1) is greatly increased by the same relative amount (**Figure 1C**). Instead of having two equally-sized effects on the components of $\hat{\beta}$ corresponding to T5 and T6, the same can be achieved in tascCODA with only one parameter by placing an effect on the internal node N3. Through **Eq. 4**, this effect is propagated to the leaves T5 and T6 in β in order to model the population.

While this aggregation step can significantly reduce the number of parameters needed to describe the changes in the data, the solution is not unique. An effect on an internal node is equivalent to effects of the same size on all its descendant leaves. Therefore, the number of nonzero entries in $\hat{\beta}$ must be controlled, raising the need for a sparse selection of the most important effects. While in the example above, the reduction of nonzero effects by using a group aggregation on node N3 clearly outweighs the loss in accuracy by assuming that features T5 and T6 behave in the same manner, this trade-off might not be as clear in real datasets. We thus also need a way to adjust the model towards selecting either more sparse and generalizing, or more detailed and less parsimonious solutions.

2.1.2 Spike-And-Slab Lasso Prior

To ease model interpretability, many statistical models provide a mechanism for obtaining sparse model solutions. In high-dimensional linear regression, this can be achieved via the lasso (Tibshirani, 1996), which adds an $\mathcal{L}_1$-penalty on the regression coefficients. In Bayesian modeling, spike-and-slab priors are a popular choice to perform automatic model selection. Recently, Ročková and George (2018), developed a connection between the two approaches in the form of the spike-and-slab lasso prior, which provides a Bayesian equivalent to penalized likelihood estimation. The spike-and-

[2]https://github.com/bio-datascience/tascCODA.

[3]https://zenodo.org/record/5302136.

slab lasso prior describes each component of $\hat{\beta}_{l,k}$ as a mixture of two double-exponential priors with different rates $\lambda_{0,l,k}$, $\lambda_{1,l,k}$ and a shared mixture coefficient θ:

$$\hat{\beta}_{l,k} = \theta\tilde{\beta}_{1,l,k} + (1-\theta)\tilde{\beta}_{0,l,k} \qquad \forall k \in [\nu], l \in [d] \quad (5)$$

$$\tilde{\beta}_{m,l,k} = \sigma_{m,l,k} * b_{m,l,k} \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (6)$$

$$\sigma_{m,l,k} \sim \mathrm{Exp}\left(\lambda_{m,l,k}^2/2\right) \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (7)$$

$$b_{m,l,k} \sim \mathcal{N}(0,1) \qquad \forall k \in [\nu], m \in \{0,1\}, l \in [d] \quad (8)$$

$$\theta \sim \mathrm{Beta}(1, 1/\nu) \quad (9)$$

This prior can be reformulated as a likelihood penalty function that represents a combination of weak penalization of larger effects by $\lambda_{1,l,k}$ and strong penalization of effects close to zero by $\lambda_{0,l,k}$, respectively (See **Supplementary Material Section 1.2**). As recommended by Ročková and George (2018), we use the non-separable version of the spike-and-slab lasso prior, which provides self-adaptivity of the sparsity level and an automatic control for multiplicity via a Beta prior on θ (Bai et al. (2020a); Scott and Berger (2010)). We further set $\lambda_{0,l,k} = 50\ \forall l, k$ to achieve a strong penalization in the "spike" part of the prior, leaving $\lambda_{1,l,k}$ as our only parameter that controls the total amount of penalty applied at larger effect values.

2.1.3 Node-Adaptive Penalization

We use a variant of the strategy proposed by Bien et al. (2021) to make the strength of the regularization penalty dependent on the corresponding node's position in the tree. We introduce the following sigmoidal scaling:

$$\lambda_{1,l,k} = 2\lambda_1 \frac{1}{1 + e^{-\phi\left(L_k/p - 0.5\right)}} \ \forall l, \quad (10)$$

where $\lambda_1 = 5$ is the default value for the penalty strength, L_k is the number of leaves that are contained in the subtree of node k, and ϕ acts as a scaling factor based on the tree structure. If $\phi = 0$, the default in tascCODA, all nodes are penalized equally with λ_1, while for $\phi < 0$, effects on nodes with larger subtrees, located closer to the root of the tree, are penalized less and are therefore more likely to be included in the model. If $\phi > 0$, a solution that comprises more diverse effects on leaf nodes will be preferred. Thus, the parameter ϕ provides a way to trade off model accuracy with the level of aggregation. We discuss the behavior of the spike-and-slab LASSO penalty and the choice of $\lambda_{0,1}$ in more detail in the **Supplementary Material**.

2.1.4 Reference Feature

Since the data at hand is compositional, model uniqueness and interpretability are only guaranteed with respect to a reference. Popular choices include picking one of the p features or the (geometric) mean over multiple or all groups (Fernandes et al., 2014). Following the scCODA model, we pick a single reference feature prior to analysis (Büttner et al., 2020). Technically, this is achieved by choosing one feature $\hat{p}$ that is set to be unchanged by all covariates. Let $\hat{\nu}$ be the set of ancestors of $\hat{p}$. By forcing $\hat{\beta}_{l,k} = 0\ \forall k \in \hat{\nu}, l \in [d]$, we ensure that the reference is not influenced by the covariates through any of its ancestor nodes. If no suitable reference feature is known a priori, tascCODA provides an automatic way of selecting the feature with minimal dispersion across all samples among the features that are present in at least a share of samples t (default $t = 0.95$; this value can be lowered if no suitable feature exists).

$$\hat{p} = \arg\min_{j=1,\ldots,p} \mathrm{Disp}\left(Y'_{\cdot,j}\right) \ s.t.\ |i\colon Y_{i,j} > 0|/n \geq t$$

The restriction to large presence avoids choosing a rare feature as the reference where small changes in terms of counts lead to large relative deviations. The least-dispersion approach is aimed at reducing the bias introduced by the choice of reference. **Eqs. 1–9** together with the reference feature yields the tascCODA model (**Figure 1D**):

$$\begin{aligned}
Y_i &\sim \mathrm{DirMult}\left(\bar{Y}_i, \mathbf{a}(X)_i\right) \\
\log\left(\mathbf{a}(X)\right)_i &= \boldsymbol{\alpha} + X_{i,\cdot}\beta \\
\alpha_j &\sim \mathcal{N}(0,10) && \forall j \in [p] \\
\beta &= \hat{\beta}A^T \\
\hat{\beta}_{l,k} &= 0 && \forall k \in \hat{\nu}, l \in [d] \\
\hat{\beta}_{l,k} &= \theta\tilde{\beta}_{1,l,k} + (1-\theta)\tilde{\beta}_{0,l,k} && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in [d] \\
\tilde{\beta}_{m,l,k} &= \sigma_{m,l,k} * b_{m,l,k} && \forall k \in \{[\nu]\backslash\hat{\nu}\}, m \in \{0,1\}, l \in [d] \\
\sigma_{m,l,k} &\sim \mathrm{Exp}\left(\lambda_{m,l,k}^2/2\right) && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in \{0,1\}, l \in [d] \\
b_{m,l,k} &\sim \mathcal{N}(0,1) && \forall k \in \{[\nu]\backslash\hat{\nu}\}, l \in \{0,1\}, l \in [d] \\
\theta &\sim \mathrm{Beta}\left(1, \frac{1}{|\{[\nu]\backslash\hat{\nu}\}|}\right)
\end{aligned}$$

with the default choices of $\lambda_{0,l,k} = 50$ and $\lambda_{1,l,k}$ set according to (10) with hyperparameters ϕ and $\lambda_1 = 5$ (**Supplementary Material Section 1.2**).

2.2 Computational Aspects

Before performing Bayesian inference with the tascCODA model, several data preprocessing steps are applied. Singular nodes, i.e., internal nodes that have only one child node, are removed from the tree, since their effect only propagates to one node and is therefore redundant. We also add a small pseudo-count of 0.5 to all zero entries of Y to minimize the frequency of numerical instabilities in our tests. Finally, we recommend normalizing all covariates to a common scale before applying tascCODA to avoid biasing the model selection process toward the covariate with the largest range of values.

Because tascCODA is a hierarchical Bayesian model, we use Hamiltonian Monte Carlo sampling (Betancourt and Girolami, 2015) for posterior inference, implemented through the tensorflow (Abadi et al., 2016) and tensorflow-probability (Dillon et al., 2017) libraries for Python, solving the gradient in each step via automatic differentiation. By default, tascCODA uses a leapfrog integrator with Dual-averaging step size adaptation (Nesterov, 2009) and 10 leapfrog steps per iteration, sampling a chain of 20,000 posterior realizations and discarding the first 5,000 iterations as burn-in, which was also the setting for all applications in this article, unless explicitly stated otherwise. As an alternative, No-U-turn sampling (Homan and

Gelman, 2014) is available for use with tascCODA. The initial states for all α_j and $b_{m,l,k}$ are randomly sampled from a standard normal distribution. All $\sigma_{m,l,k}$ and θ values are initialized at 1 and 0.5, respectively.

To determine the credible effects of covariates on nodes from the chain of posterior samples, we calculate the threshold of practical significance δ_k, introduced by Ročková and George (2018), for each node:

$$\delta_k = \frac{1}{\lambda_0 - \lambda_{1,k} \log\left(\frac{1}{p^*_{\theta,k}(0)} - 1\right)} \tag{11}$$

$$p^*_{\theta,k}(\beta) = \frac{\theta^* \frac{\lambda_{1,k}}{2} e^{-\lambda_{1,k}|\beta|}}{\theta^* \frac{\lambda_{1,k}}{2} e^{-\lambda_{1,k}|\beta|} + (1-\theta^*)\frac{\lambda_0}{2} e^{-\lambda_0|\beta|}} \tag{12}$$

Here, $\theta^\star$ is the posterior median of θ. More details on δ are available in the **Supplementary Material**. We compare the posterior median effects $\hat{\beta}^*_{l,k}$ to the corresponding δ_k and select all effects where $|\hat{\beta}^*_{l,k}| > \delta_k$ as credible, otherwise they will be set to 0, resulting in $\hat{\beta}^{(C)}$, the matrix with only credible effects,

$$\hat{\beta}^{(C)}_{l,k} = \begin{cases} \hat{\beta}^*_{l,k} & \text{if } |\hat{\beta}^*_{l,k}| > \delta_k \\ 0 & \text{else.} \end{cases} \tag{13}$$

In most applications, the nonzero entries of $\hat{\beta}^{(C)}$ are of primary interest, which directly show how the covariates influence sets of features defined by the tree structure. Their sign indicates whether the effect corresponds to an increase ($\hat{\beta}^{(C)}_{l,k} > 0$) or a decrease ($\hat{\beta}^{(C)}_{l,k} < 0$). Due to the compositional data properties introduced by the Dirichlet-Multinomial, its expectation

$$\mathrm{E}\left[Y_i \sim \mathrm{DirMult}\left(\bar{Y}_i, \mathbf{a}(\mathbf{x})_i\right)\right] = \bar{Y}_i \frac{\mathbf{a}(\mathbf{x})_i)}{\sum_{j=1}^{p} \mathbf{a}(\mathbf{x})_i)_j} \tag{14}$$

can not be separated by the individual features. Because the shifts in $\mathrm{E}[Y_i]$ caused by effects $\hat{\beta}$ are dependent on the total sum $\sum_{j=1}^{p} e^{\alpha_j + X(\hat{\beta}A^T)_j}$ through **Eqs. 2**, **4**, **14**, a credible effect on any feature or aggregation has an impact on the posterior mean counts of all features, i.e. a relative increase in one feature will also induce a decrease of all other features (Gloor et al., 2017). Therefore, a quantitative interpretation of effect sizes is only possible in a limited sense. Within the same model, larger changes will correspond to larger absolute values $|\hat{\beta}_{l,k}|$, but they are not comparable across multiple runs of tascCODA.

In the context of differential abundance testing, we can additionally obtain the set of differentially abundant features D by multiplying $\hat{\beta}^{(C)}$ with A^T, and get

$$D = \left\{(l, j) \in [d] \times [p] : \left(\hat{\beta}^{(C)}_{l,k} A^T\right)_j \neq 0\right\} \tag{15}$$

as the set of features that are part of at least one credible effect.

A Python package for tascCODA is available at https://github.com/bio-datascience/tascCODA. Building upon the scCODA package, the software provides methods to seamlessly integrate scRNA-seq data from scanpy (Wolf et al., 2018) or microbial population data via pandas (McKinney, 2010). The package also allows to perform differential abundance testing with tascCODA and visualize tascCODA's results through tree plots from the toytree package. All results were obtained using Python 3.8 with tensorflow = 2.5.0 (Abadi et al. (2016)), tensorflow-probability = 0.13 (Dillon et al. (2017)), arviz = 0.11 (Kumar et al. (2019)), numpy = 1.19.5, scanpy = 1.8.1 (Wolf et al. (2018)), toytree = 2.0.1, and sccoda = 0.1.4 (Büttner et al. (2020)).

3 RESULTS

3.1 Simulation Studies

3.1.1 Model Comparison

To test the performance of tascCODA in a differential abundance testing scenario, we generated compositional datasets with an underlying tree structure and compared how well several models could detect the changes introduced by a binary covariate. For compositional models that do not account for the tree structure, we used the state-of-the art methods ANCOM-BC (Lin and Peddada (2020)), ANCOM (Mandal et al. (2015)), and ALDEx2 (Fernandes et al. (2014)) from the field of microbiome data analysis, as well as scCODA (Büttner et al., 2020) from scRNA-seq analysis. Based on the recommendations by Aitchison (1982), we also analyzed the data with the additive log-ratio (ALR) transformation in combination with t- or Wilcoxon rank-sum tests. We also included the recent adaANCOM (Zhou C. et al., 2021), a differential abundance testing method that accounts for the tree structure. Furthermore, we applied tascCODA with different values for the aggregation parameter, ϕ = (−10, −5, −1, 0, 1, 5, 10), setting $\lambda_1 = 5$.

We first defined four different data sizes p = (10, 30, 50, 100) and randomly generated a multifurcating tree with depth five for each value of p. We then chose three nodes (one internal on the level directly above the leaves, two leaves) from each tree, whose child leaves, denoted by p', are set to be differentially abundant under a binary (control-treatment) condition (**Supplementary Figures S2–S5**). Similar to Wadsworth et al. (2017), we generated $n = n_0 + n_1$ compositional data samples from two groups of equal size $n_0 = n_1$ = (5, 20, 30, 50). Each sample Y_i is a realization of a Dirichlet-Multinomial distribution with a total sum of $\bar{Y}_i$ = 10, 000 and a parameter vector $\gamma^\star$. For extra dispersion in the data, we set $\gamma^*_i = \frac{\gamma_i}{\sum_j \gamma_j} \frac{1-\psi}{\psi}$ with $\psi = 0.002$. The parameters for the first (control) group were generated via $\gamma_{0,i} = \exp(\alpha_i)$; α_i ~Unif(−2, 2). In the second (treatment) group, we added an effect β = (0.3, 0.5, 0.7, 0.9) to the components in p': $\gamma_{1,i} = \exp(\alpha_i + \beta \mathbb{I}_{(i \in p')})$. For each parameter combination (p, n_0, β), we randomly generated 20 replicates, resulting in a total of 1280 datasets.

Since the adaANCOM method assumes a bifurcating tree structure, we transformed each tree node to a series of bifurcating splits via the *multi2di* and *collapse.singles* methods from the *ape* package for R (Paradis et al. (2004)) before applying the method. For the methods that require a reference category (ALR, scCODA, tascCODA, ALDEx2), we used the last component, which was always designed to be unaffected by

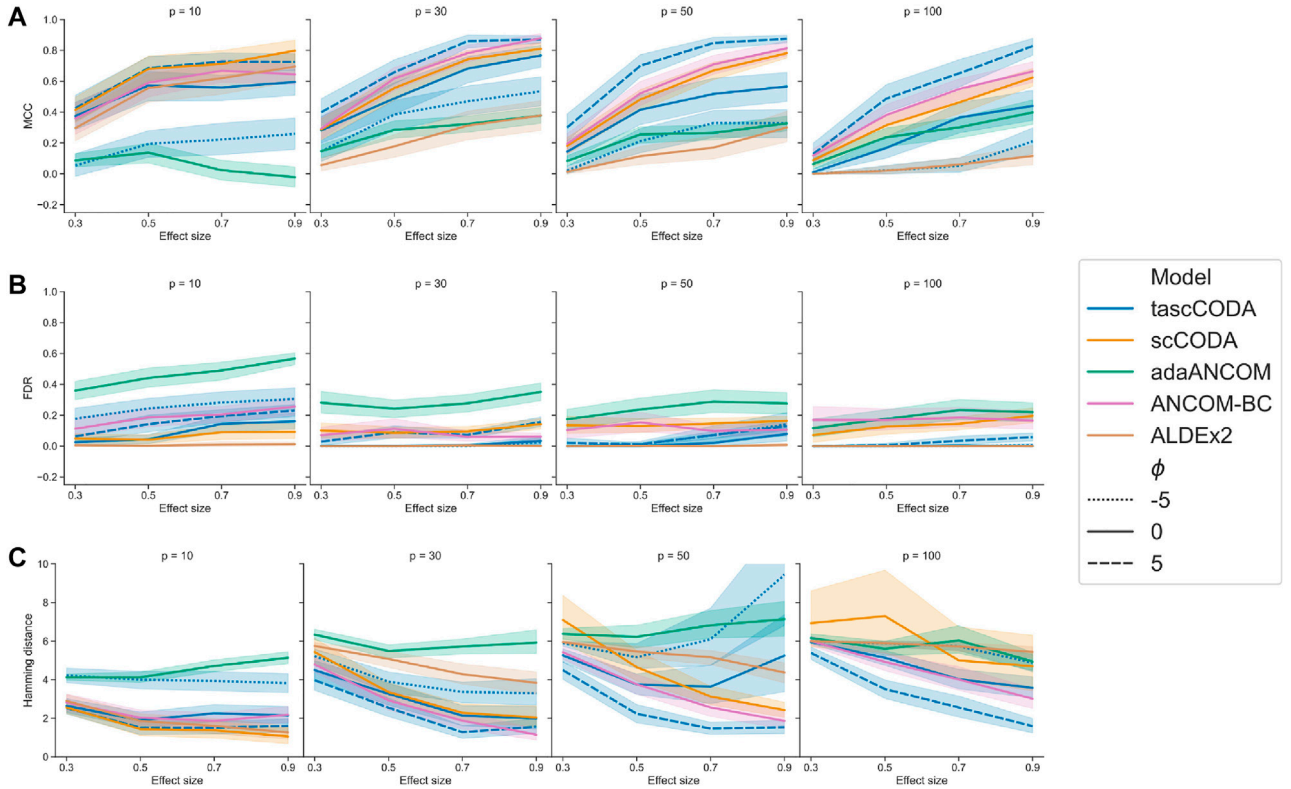


FIGURE 2 | Performance comparison of tascCODA and other methods on simulated data with one binary covariate (differential abundance testing). Plots are grouped by the number of simulated components p and the effect size β. For tascCODA, different values of ϕ were tested (dashed blue lines). The areas around each line represent the standard deviation. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

the condition, as the reference. After applying each method to a dataset, we corrected the resulting p-values by the Benjamini-Hochberg procedure, where applicable, except for ANCOM-BC, where we used the recommended Holm correction of p-values, and determined the significant results at an expected FDR level of 0.05. The Bayesian methods scCODA and tascCODA do not produce p-values and identify credible effects as previously described.

For an overall indicator of how well the different methods could determine differentially abundant features, we considered Matthews correlation coefficient (**Figure 2A**). Here, adaANCOM showed poor performance especially on small datasets, while ALDEx2 struggled when p was larger. Only scCODA and ANCOM-BC performed well in comparison for all data and effect sizes. For tascCODA, varying the aggregation level ϕ had a strong influence on the performance. With larger values of ϕ, tascCODA prefers less generalizing effects, resulting in a more detailed solution and larger MCC. At a high resolution level ($\phi = 5$), tascCODA was on par with or even better than scCODA and ANCOM-BC, showing almost no sensitivity to the size of the dataset. Because the trees in our simulation contained only effects on leaf nodes or the level directly above, preferring generalizing effects ($\phi = -5$) resulted in worse performance, while the unbiased case of $\phi = 0$ gave slightly worse results than scCODA and ANCOM-BC. All methods shown in **Figure 2B** except adaANCOM controlled the FDR reasonably well, although ANCOM-BC and scCODA could not always hold the nominal level of 0.05. Only ALDEx2, which is known to be very conservative (Hawinkel et al., 2019; Büttner et al., 2020), produced almost no false positives, at the cost of larger type 2 error. tascCODA had a slightly inflated FDR (< 0.25) for smaller values of ϕ in some cases, which became more apparent when analyzing the ability of each method to exactly recover the true effects (**Figure 2C**). Increasing the effect size resulted in a reduced Hamming distance between the ground truth and tascCODA with $\phi = 5$, which consistently outperformed all other models. tascCODA in the misspecified setting $\phi = -5$ showed an inflated Hamming distance, especially for $p = 30$. This is, however, expected since tascCODA is forced to infer small-sized effects at the top level, resulting in many falsely detected features and thus a large deviation from the true sparse solution. In practice, this highlights the need to perform cross-validation over different levels of ϕ to reduce false discoveries due to misspecification. We further found that ANCOM detected many false positives in all of our simulations, while the ALR-based methods were similarly

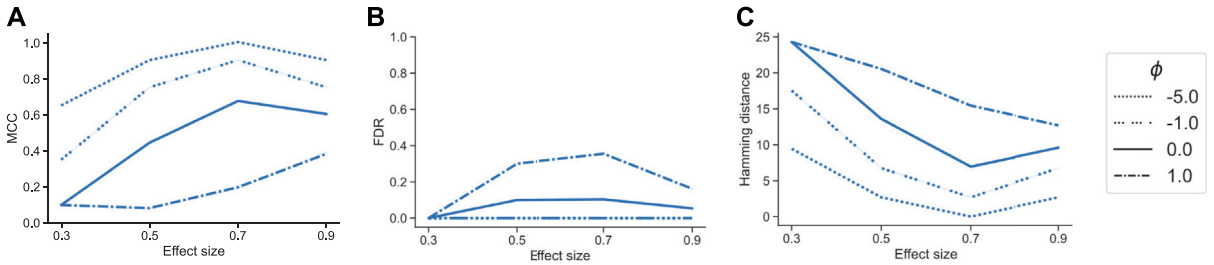


FIGURE 3 | Performance comparison of different bias settings for tascCODA on simulated data with the effect being located near the root of the tree, depending on effect size. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

conservative as ALDEx2 (**Supplementary Figures S8–S10**). Increasing the sample size generally improved the recovery performance of all methods except for tascCODA with misspecified ϕ (**Supplementary Figure S10**).

3.1.2 Effect Detection at High Tree Levels

In the next benchmark scenario, we evaluated the effect of the tuning parameter ϕ in tascCODA to detect effects on larger groups of features through aggregation at higher levels of the tree. To this end, we considered the $p = 30$ setting with the tree structure from **Supplementary Figure S5**, and defined an effect on a node near the root, influencing almost all features (**Supplementary Figure S6**). We simulated datasets in the same manner as for the previous benchmark, with $n = 10$, $\beta = (0.3, 0.5, 0.7, 0.9)$, and 20 replicates per effect size. We then compared tascCODA with different levels of ϕ using the same performance metrics as before.

With a correctly specified parametrization $\phi < 0$, favoring effects near the root, tascCODA recovered almost all relevant effects, as indicated by a small Hamming distance and high MCC, without producing false positive results (**Figure 3**). With increasing ϕ, however, tascCODA favors effects on the leaves, thus entering the misspecified regime. As predicted, tascCODA was able to only recover a small portion of the true effects, while producing more false positive results. This highlights tascCODA's ability to consistently uncover effects on larger groups of features which would be missed when not taking into account tree information.

3.1.3 Simulation With Multiple Covariates

In our third benchmark scenario, we simulated data with two covariates to showcase how tascCODA is able to distinguish effects from two different sources. Taking the tree from the method comparison study with $p = 30$ (**Supplementary Figure S3**), we first defined a binary covariate x_0 with effect sizes $\beta_0 = (0.3, 0.5, 0.7, 0.9)$ as before, and $n = 10$ samples per group. We also included a second covariate $x_1 \sim Unif(0, 1)$ with effect size $\beta_1 = 3$ that affects node 39 and therefore features 13–23 in all samples. For each effect size, we simulated 10 datasets and applied tascCODA with $\phi = (-5, 0, 5)$ and two different design matrices X. For the first design matrix, we used only x_0, while the second design matrix contained both x_0 and x_1 as covariates. We compared how well both configurations could recover the effects introduced by x_0 in terms of MCC, FDR, and Hamming distance to the ground truth.

Ignoring x_1 in the model design resulted in an overall worse performance of tascCODA for all metrics, all effect sizes for x_0, and all values of ϕ (**Figure 4**). In every case it proved beneficial to include the second covariate in the model, resulting in almost no false positive detections of changes caused by the first covariate. Further, the two-covariate model achieved an MCC and Hamming distance that were similar to our simulations where only one covariate acted on the data (**Figure 2**). This proves that tascCODA is able to reliably identify the influence of multiple covariates on the count data.

3.2 Experimental Data Applications

3.2.1 Single-cell Sequencing Analysis of Ulcerative Colitis in Humans

Ulcerative colitis is one of the most common manifestations of inflammatory bowel disease. The disease alternates between periods of symptomatic flares and remissions. The flares are due to the surge of an inflammatory reaction in the colon, causing superficial to profound ulcerations, which manifests with bloody stool, diarrhea and abdominal pain. The patients will thus have part of their colon referred to as "inflamed", while colonic tissue still seemingly intact will be called "non-inflamed". To show how tascCODA can be applied to cell population data from scRNA-seq experiments, we used data collected by Smillie et al. (2019) from a study of the colonic epithelium on ulcerative colitis (UC). In the study, a total of 133 samples from 12 healthy donors, as well as inflamed and non-inflamed tissue from 18 patients with UC, were obtained via single-cell RNA-sequencing, divided into epithelial samples and samples from the Lamina Propria (**Supplementary Data 1.3.1**).

We applied tascCODA to six different subsets of the data, comparing two of the three health conditions in one type of tissue at a time, and then compared our findings with the results of scCODA and the Dirichlet regression model used by Smillie et al. (2019), implemented in the *DirichletReg* package for R (Maier (2014)). For tascCODA and scCODA, we used the automatically determined reference cell types, which are identical for both models in all cases, and applied scCODA

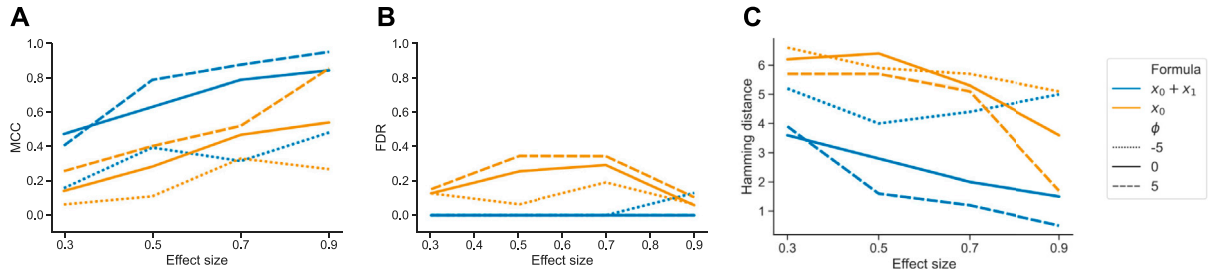


FIGURE 4 | Performance comparison for tascCODA on simulated data with two covariates. The setups including both or only one covariate in the model are shown as $x_0 + x_1$ and x_0, respectively. Simulations were evaluated for different effect sizes and aggregation levels ϕ. Performance measured by **(A)** Matthews correlation coefficient (MCC). **(B)** False discovery rate (FDR) **(C)** Hamming distance between ground truth and determined effects.

with an FDR level of 0.05. In the Dirichlet regression model, we adjusted the *p*-values by the Benjamini-Hochberg procedure, and selected differentially abundant cell types at a level of 0.05.

The cell lineage tree inferred from Smillie et al. (2019) is divided into epithelial, stromal and immune cells at the top level (**Figure 5**). While the biopsies from the Epithelium contain mostly epithelial cells, and samples from the Lamina Propria consist of cells mostly from the other two lineages, both groups also include considerable amounts of cells from the other major lineages. We first compared scCODA and Dirichlet regression, which both do not take the tree structure into account, to tascCODA with $\phi = 5$ (**Figure 6**), thus preferring a detailed solution with effects mainly located on leaf nodes, which approaches the leaf-only solutions of the other two methods. In this setting, tascCODA, scCODA and Dirichlet regression all determined mostly epithelial cells to shift in abundance between pairwise comparisons of healthy, non-inflamed, and inflamed tissue samples from the intestinal Epithelium (**Figure 6A**), and most changes in the Lamina Propria to be among stromal and immune cells (**Figure 6B**). When propagating the node effects of tascCODA with $\phi = 5$ to the leafs via **Eq. 15**, the differentially abundant cell types determined by tascCODA, scCODA, and Dirichlet regression were largely identical (**Figure 6**).

To further investigate the predictive and sparsity-inducing powers of tascCODA, we performed out-of-sample prediction with the results obtained from tascCODA and scCODA on 5-fold cross validation splits of each of the six data subsets. For both models, we determined cell type-specific effect vectors $\beta^\star$ (tascCODA: $\beta^* = A\hat{\beta}_j^{(C)}$, as in **Eq. 15**; scCODA: Model output) as well as the posterior mean of the base composition $\alpha^\star$ on the training splits, and used them to predict cell counts for each health status label X_l in the corresponding test split as $\hat{y}_{j,l} = \frac{e^{\alpha_j^* X_l \beta_j^*}}{\sum_{j=1}^{p} e^{\alpha_j^* X_l \beta_j^*}} \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \bar{Y}_i$. We measured the predictive power of tascCODA and scCODA as the mean squared logarithmic error (MSLE) between the actual and predicted cell counts, and sparsity as the average number of nonzero effects over all five splits (**Table 1**). For small ϕ, tascCODA determined very few or no credible effects, while the MSLE was usually slightly higher than the MSLE from scCODA. In unbiased setting $\phi = 0$, tascCODA found credible effects in three scenarios, which considerably reduced the MSLE. With a small bias towards the leaves ($\phi = 1$), tascCODA even outperformed scCODA in terms of MSLE in one case, while for $\phi = 5$, tascCODA achieved a lower MSLE and similar number of credible effects in three scenarios, and a lower number of credible effects and similar MSLE in the other three scenarios. We observed a curious result when comparing non-inflamed and inflamed epithelial samples. Here, the MSLE increased with rising ϕ, indicating that the mean model over all samples described the data better than trying to determine variation between the two groups. This confirms the intuition that the aggregation bias ϕ in tascCODA acts as a trade-off between generalization level and prediction accuracy. For smaller ϕ, tascCODA will select fewer, more general effects, which might miss subtle changes at a lower level of the lineage tree, while with increasing ϕ, tascCODA's results will approach the ones discovered without taking tree aggregation into account.

For a more detailed comparison between tascCODA and scCODA, we compared healthy to non-inflamed biopsies of control and UC patients. When choosing $\phi = 5$, thus biasing tascCODA towards the leaf nodes, tascCODA detected the differences in cell composition in the Epithelium as changes in abundance of the same 3 cell types as scCODA (**Figure 5A**). In the Lamina Propria, tascCODA detected credible changes on six different groups of cell types, including T and B cells, which were previously linked to UC (Holmén et al. (2006); Smillie et al. (2019)), as well as eight single cell types (**Figure 5B**). Notably, tascCODA amplified the decrease of Plasma B-cells induced by the group effect on B-cells by an additional negative effect on the cell type level. A strong decrease of Plasma cells was also confirmed by Smillie et al. (2019) through FACS stainings. Importantly, tascCODA described the data with only 14 nonzero effects, whereas with scCODA, 21 credible effects were produced.

As a contrast, we also examined the unbiased setting with $\phi = 0$, treating all nodes equally. Here, the cell type-specific changes in the Epithelium were not picked up anymore by tascCODA (**Figure 5C**). In the Lamina Propria, only seven effects, almost all on groups of cell types, were detected by tascCODA

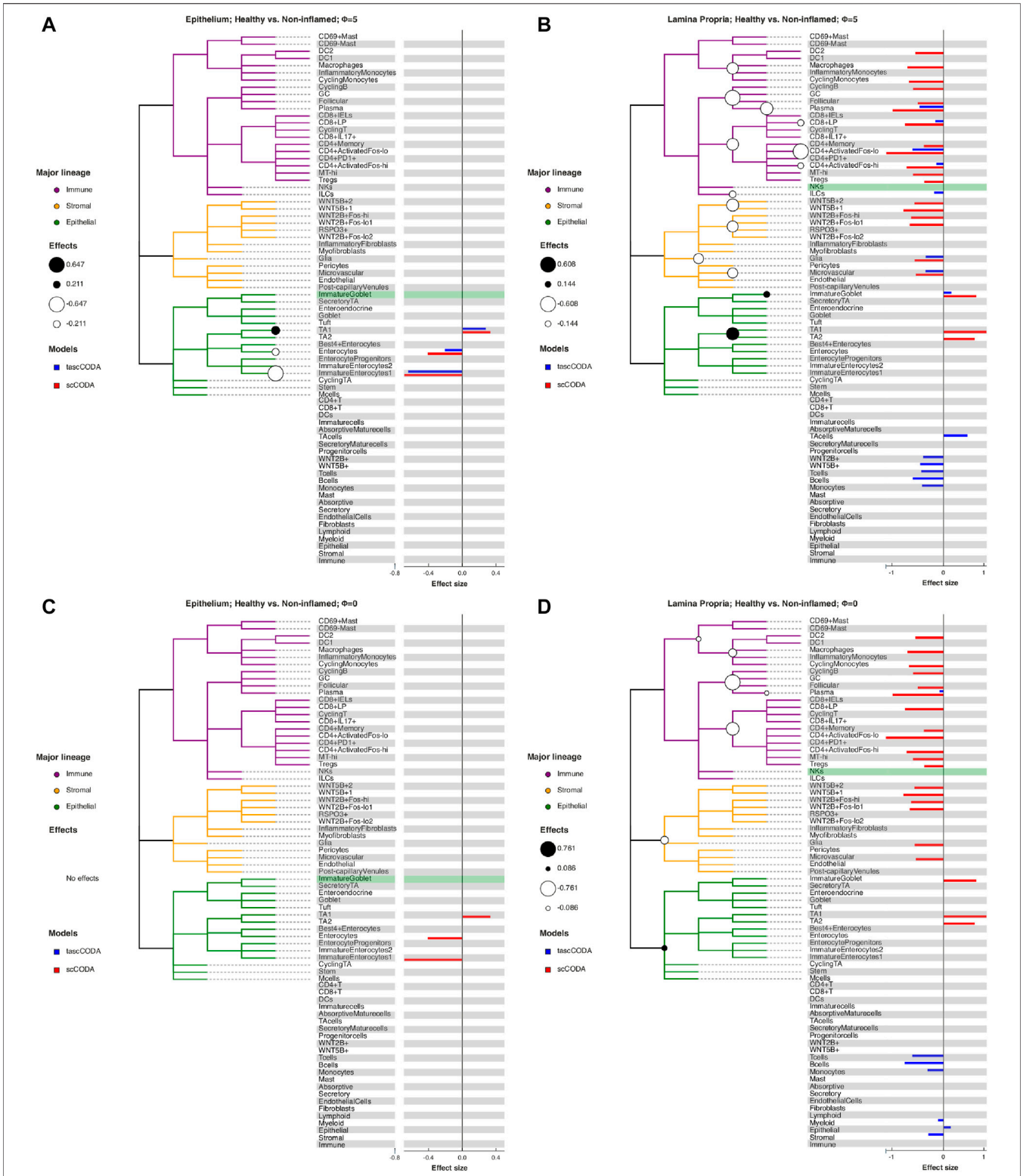


FIGURE 5 | Behavior of tascCODA on scRNA-seq data for different values of ϕ. All plots show the comparison of healthy control samples to non-inflamed tissue samples of UC patients in the data from Smillie et al. (2019). White and black circles on the cell lineage tree show the effects found by tascCODA, which are also shown as blue bars on the right side of each plot. The bars below the tree depict effects on internal nodes, with lower positions in the diagram corresponding to nodes closer to the root. For comparison, the red bars indicate effects found by scCODA, which only operates on the tips of the tree. The green-shaded area shows the reference cell type that was used for both models. **(A)** When $\phi = 5$, tascCODA prefers placing effects near the tips of the tree and finds the exact same solution as scCODA for the Epithelium data. **(B)** In the Lamina Propria, tascCODA places some effects on internal nodes, resulting in a sparser solution than the one obtained by scCODA (14 vs. 21 credible effects). **(C)** When $\phi = 0$, tascCODA finds no credible effects in samples from the Epithelium, and **(D)** only seven effects are necessary to summarize the large number of effects found by scCODA when looking at samples from the Lamina Propria.

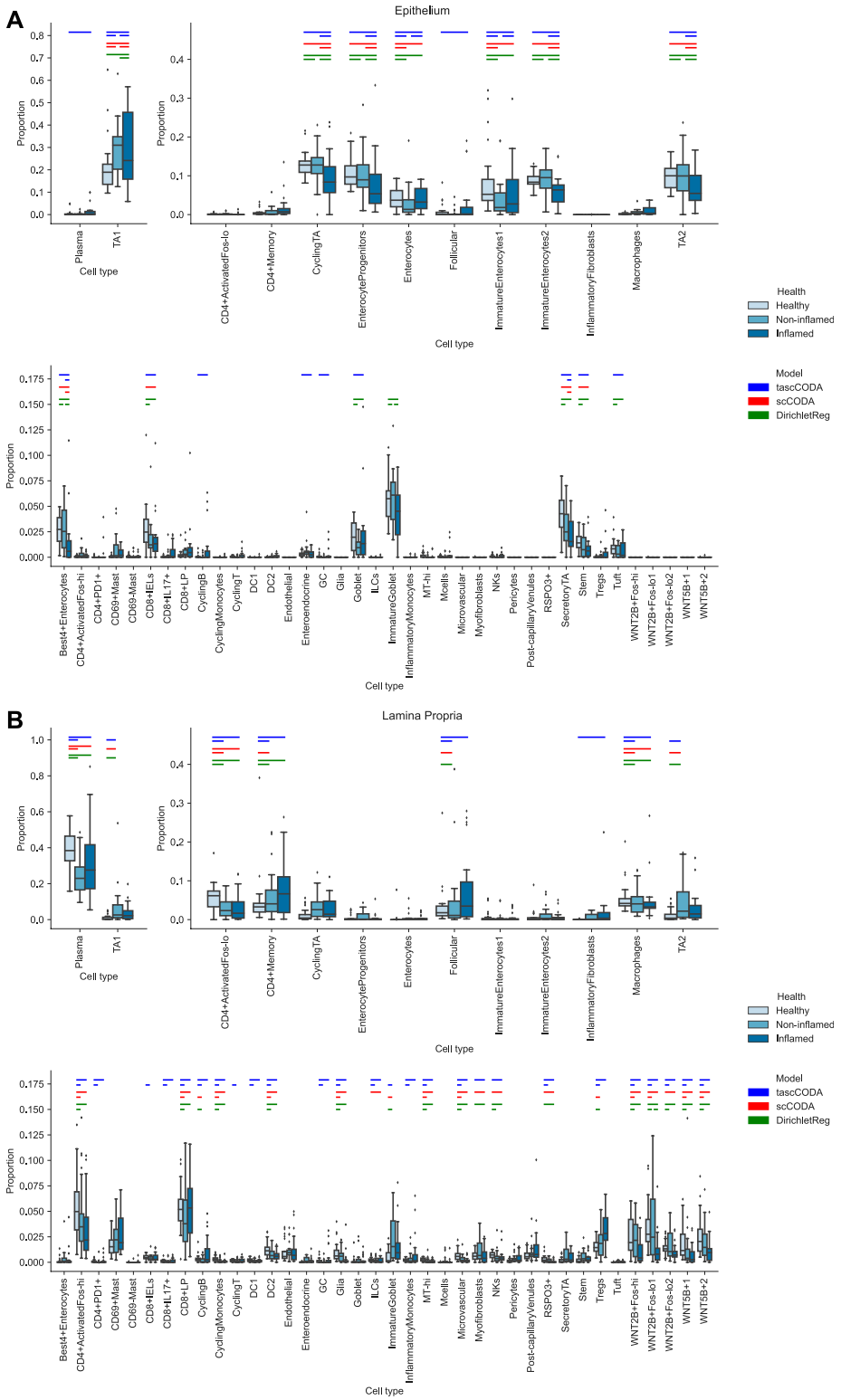


FIGURE 6 | Comparison of differentially abundant cell types found by tascCODA (blue, $\phi = 5$), scCODA (red, FDR = 0.05), and Dirichlet regression (green, adjusted $p_{adj} < 0.05$) between biopsies of healthy, non-inflamed and inflamed tissue. Colored bars for each method indicate that a credible change was found. **(A)** Among samples from the intestinal epithelium, tascCODA and Dirichlet regression detect effects on lowly abundant epithelial cell types (Tuft, Goblet, Enteroendocrine) that were not detected by scCODA. **(B)** In the Lamina Propria, only tascCODA detects a number of effects on some of the T and B cell types.

TABLE 1 | Mean squared logarithmic error (MSLE) and number of selected effects over five cross-validation splits for tascCODA with different parametrizations ϕ and scCODA. Abbreviations for scenarios: Healthy (H), Non-inflamed (N), and Inflamed (I). With increasing ϕ, tascCODA selects more effects and on average improves its predictive power. At $\phi = 5$, tascCODA has equal or lower MSLE than scCODA and a similar number of selected effects.

	Model	tascCODA					scCODA
Scenario	ϕ	−5	−1	0	1	5	-
Epithelium - H vs. N	MSLE	142.22	142.16	142.18	138.56	134.36	134.96
	Effects	0.0	0.0	0.0	1.2	3.2	2.4
Epithelium - H vs. I	MSLE	167.46	163.60	160.68	158.06	154.64	154.44
	Effects	0.0	1.6	2.6	3.2	8.2	10.8
Epithelium - N vs. I	MSLE	173.94	174.10	174.10	175.86	177.26	174.78
	Effects	0.0	0.0	0.0	0.2	3.6	5.2
LP - H vs. N	MSLE	162.76	157.62	155.16	152.80	149.58	154.02
	Effects	0.4	1.8	3.0	6.2	16.0	14.4
LP - H vs. I	MSLE	188.58	182.96	178.88	176.02	173.32	173.40
	Effects	0.0	1.8	4.8	7.8	17.8	17.4
LP - N vs. I	MSLE	219.72	219.70	219.66	219.68	216.76	218.62
	Effects	0.0	0.0	0.0	0.0	1.4	0.4

(**Figure 5D**). Again, B and T cells were found as the cell lineages that undergo the largest change between healthy and non-inflamed UC biopsies. When testing healthy versus inflamed, and non-inflamed versus inflamed biopsies, tascCODA also detected more detailed results when $\phi = 5$, and found fewer, more generalizing effects with $\phi = 0$ (**Supplementary Figures S11, S12**; **Supplementary Tables S1–S3**).

3.2.2 Analysis of the Human Gut Microbiome Under Irritable Bowel Syndrome

We next considered a microbiome data example and focused on another chronic disorder of the human gut, the Irritable Bowel Syndrome (IBS). IBS is a functional bowel disorder characterized by frequent abdominal pain, alteration of stool morphology and/or frequency, with the absence of other gastrointestinal diseases (i.e. colorectal cancer, inflammatory bowel disease). It is estimated that about 10% of the general population experience symptoms that can be classified as a subtype of Irritable Bowel Syndrome, which include IBS-C (constipation), IBS-D (diarrhea), IBS-M (mixed), or unspecified IBS (Ford et al. (2017)). While the exact sources of the disease can be manifold, it has been hypothesized that the gastroenterological symptoms may be caused by a disturbed composition of the gut microbiome (Duan et al. (2019); Ford et al. (2017)).

In particular, we analyzed 16S rRNA sequencing data of stool samples collected from IBS patients and healthy controls, which were obtained by Labus et al. (2017). The dataset consists of $n = 52$ samples, with 23 healthy controls, and 29 IBS patients separated into 11 subjects with constipation (IBS-C), 10 subjects with diarrhea (IBS-D), 6 subjects with mixed symptoms (IBS-M), and 2 subjects with unspecified symptoms. Further, metadata information about age, sex and BMI of most subjects is available. We re-processed the raw 16S rRNA sequences with DADA2, version 1.21.0 (Callahan et al. (2016)) and did taxonomic assignment via the Silva database, version 138.1 (Quast et al. (2013); Yilmaz et al. (2014)), yielding a final count table with 709 ASVs along with a taxonomic tree (**Supplementary Data 1.3.2**). This data was then aggregated at the genus level, resulting in a total of $p = 91$ known genera.

We applied tascCODA to the genus-level data, comparing healthy and IBS subjects. To showcase the flexibility of tascCODA, we analyzed the data with different covariate setups, by including the other available metadata variables. As a reference genus for scCODA and tascCODA, we chose *Alistipes*, since it is a genus with relatively high presence and rather low dispersion. For all analyses on this dataset, we decreased the mean shrinkage in tascCODA to $\lambda_1 = 1$, allowing us to find more subtle effects.

We first used tascCODA to analyze the differences in the gut microbial composition between healthy controls and IBS patients (**Figure 7**, **Supplementary Table S4**). Favoring generalization with $\phi = -5$, we found only a small decrease of the phylum Firmicutes (**Figure 7A**). In the unbiased setting ($\phi = 0$), the previous effect on the phylum level was substantiated to the Oscillospirales order. Additionally, decreases of the *Parabacteroides* and *Bacteroides* genera are found (**Figure 7B**). Setting $\phi = 5$, thus favoring detailed results, we discovered a decrease of the Ruminococcaceae family, a subgroup of Oscillospirales, and multiple decreasing genera with the strongest effects on *Parabacteroides* and *Bacteroides* (**Figure 7C**). For comparison, we also applied scCODA (FDR = 0.1) to the same dataset, which also discovered a decrease of *Parabacteroides* and *Bacteroides*, as well as three genera in the Ruminococcaceae family. A decrease of *Parabacteroides* in a subset of IBS patients was also found by Labus et al. (2017). Also, a relative decrease of the order Bacteroidales, which includes *Parabacteroides* and *Bacteroides*, was reported by Nagel et al. (2016) and Jeffery et al. (2012). Decreasing shares of Ruminococcaceae were also connected to IBS in multiple studies (Durbán et al., 2012; Pozuelo et al., 2015).

To highlight the flexibilty of tascCODA, we next tried to discover changes in the gut microbiome related to age, BMI, gender, and IBS subtype. Before applying tascCODA, we min-max normalized the two former covariates to obtain a common scale for all covariates. We excluded three samples with missing information on BMI. We conducted every analysis three times with $\phi = -5, 0, 5$. When testing for changes related to one of age, gender, or BMI alone, tascCODA

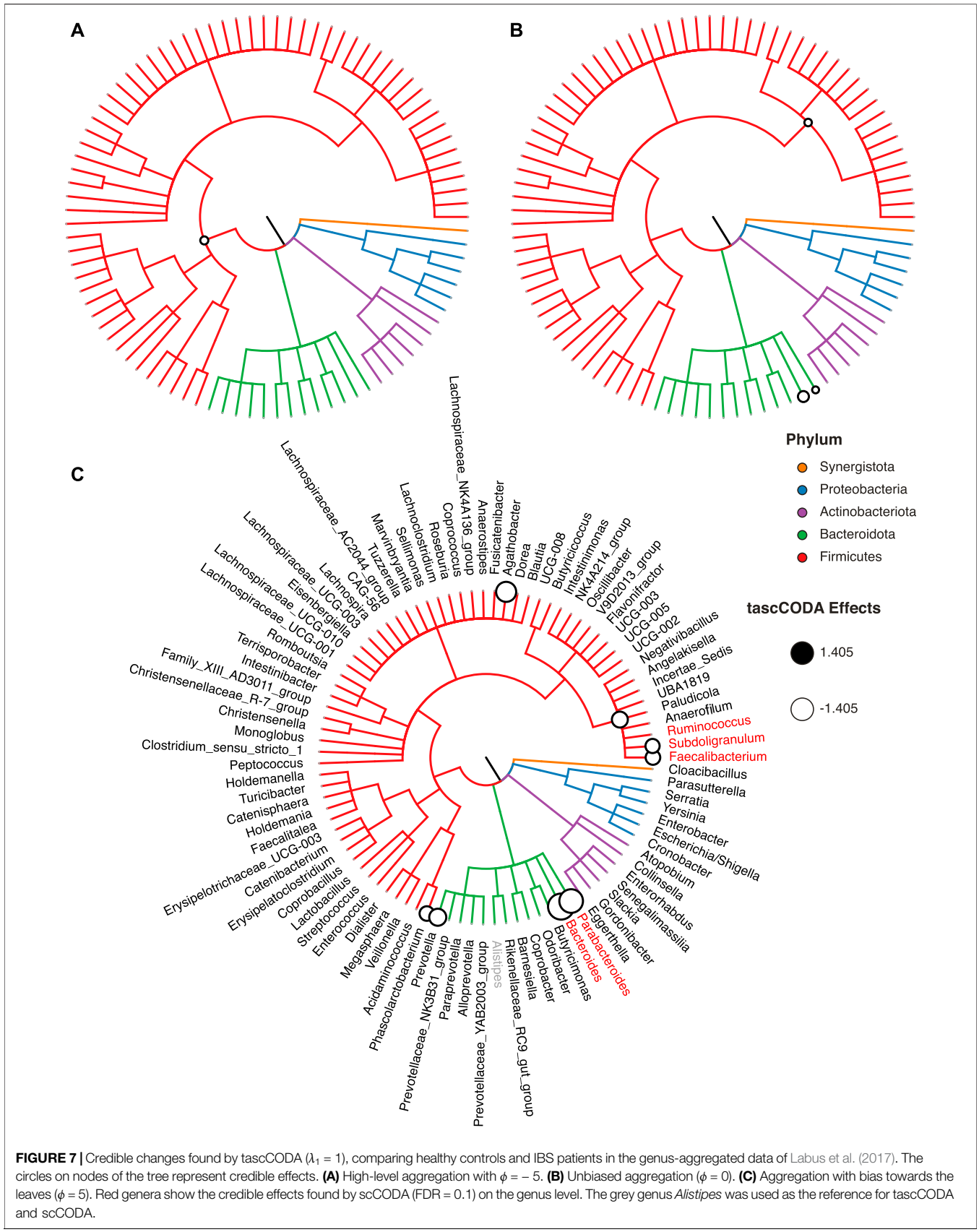


FIGURE 7 | Credible changes found by tascCODA ($\lambda_1 = 1$), comparing healthy controls and IBS patients in the genus-aggregated data of Labus et al. (2017). The circles on nodes of the tree represent credible effects. **(A)** High-level aggregation with $\phi = -5$. **(B)** Unbiased aggregation ($\phi = 0$). **(C)** Aggregation with bias towards the leaves ($\phi = 5$). Red genera show the credible effects found by scCODA (FDR = 0.1) on the genus level. The grey genus *Alistipes* was used as the reference for tascCODA and scCODA.

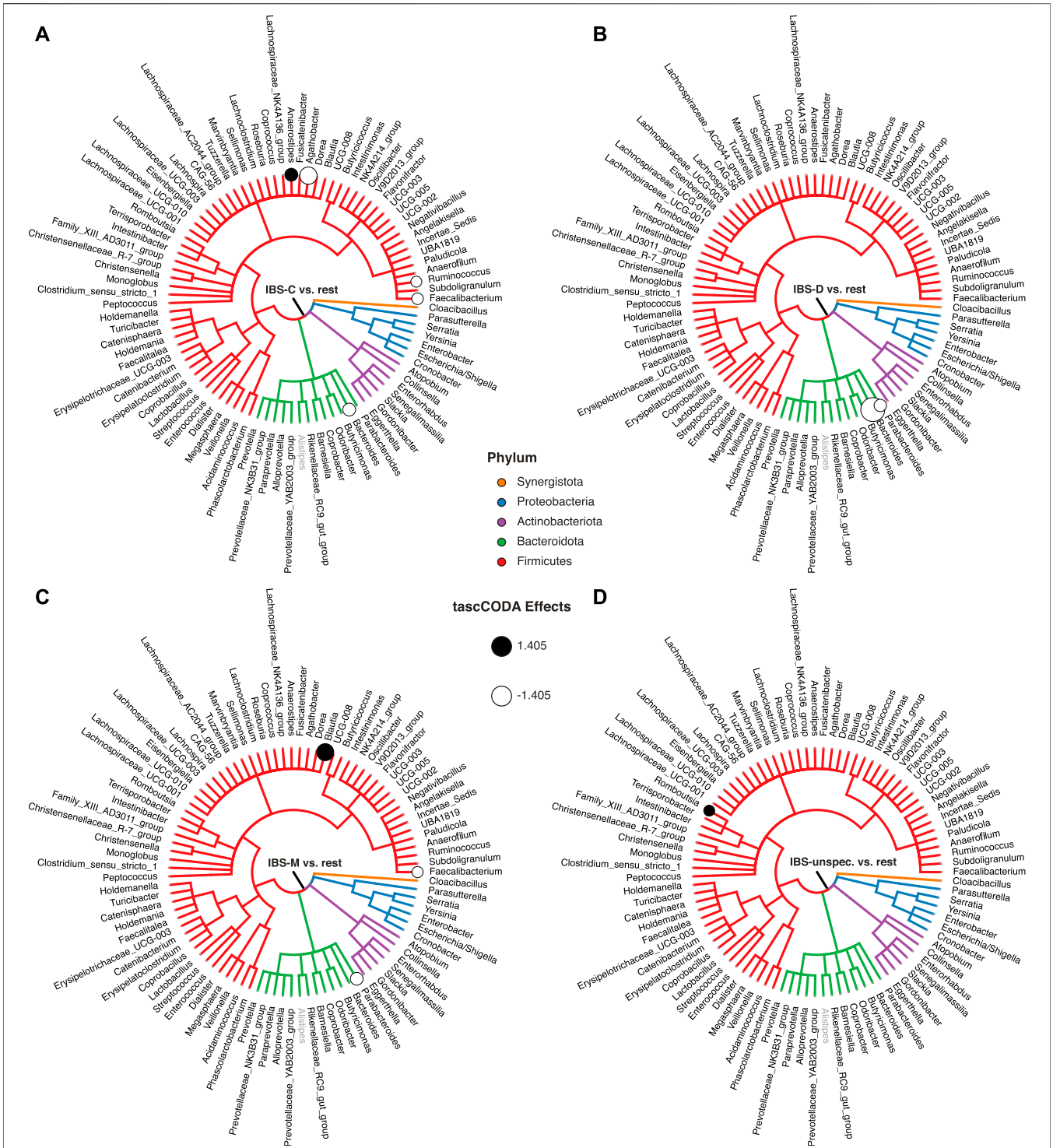


FIGURE 8 | Credible changes found by tascCODA ($\lambda_1 = 1$, $\phi = 5$), simultaneously comparing healthy controls to all IBS subtypes in the genus-aggregated data of Labus et al. (2017). The circles on nodes of the tree represent credible effects. The grey genus *Alistipes* was used as the reference for tascCODA. **(A)** IBS-C ($n = 11$). **(B)** IBS-D ($n = 10$). **(C)** IBS-M ($n = 6$). **(D)** IBS-unspecified ($n = 2$).

was not able to discover any credible differences for any aggregation bias. When testing on all four covariates together, excluding interactions, tascCODA only reported credible changes in the microbiome with respect to the IBS subtype. Finally, including all possible variables, interactions revealed that while a general negative effect was found independent of gender, male IBS-D patients had a larger depletion of *Bacteroides* than female patients.

Next, we restricted our analysis to testing for changes between the four IBS subtypes and all other samples. The results shown in **Figure 8** and **Supplementary Table S5** were obtained with $\phi = 5$. For patients experiencing constipation (IBS-C, **Figure 8A**), decreases of *Agathobacter*, *Bacteroides*, *Ruminococcus*, and *Faecalibacterium*, as well as an increase of *Anaerostipes* were found by tascCODA. Conversely, diarrhea (IBS-D, **Figure 8B**) was associated with a decrease in *Parabacteroides*, as well as a large decrease in *Bacteroides*. Patients with mixed symptoms (IBS-M, **Figure 8C**) were found to have increased numbers of *Blautia*, in addition to a decrease of *Parabacteroides* and *Faecalibacterium*, which each match with the observations related to one of the two previous conditions. Finally, only a small increase of *Romboutsia* was associated to IBS with unspecified symptoms (IBS-unspecified, **Figure 8D**).

4 DISCUSSION

Associating changes in the structure of microbial communities or cell type compositions with host or environmental covariates are commonly investigated with amplicon or single-cell RNA sequencing. With tascCODA, we have presented a fully Bayesian method to determine such compositional changes that acknowledges the hierarchical structure of the underlying microbial or cell type abundances and simultaneously accounts for the compositional nature of the data. By introducing tree-based penalization that adapts to the structure of the tree, the tascCODA model is able to accurately identify group-level changes with fewer parameters than traditional individual feature-based approaches. Thanks to a scaled variant of the spike-and-slab lasso prior (Ročková and George (2018)), we were able to obtain sparse solutions that can favor high-level aggregations or more detailed effects on a dynamic range characterized by a single scaling parameter ϕ. The tascCODA Python package seamlessly integrates into the *scanpy* environment for scRNA-seq (Wolf et al. (2018)) and allows Bayesian regression-like analyses with flexible covariate structures.

Through its ability to favor general trends or more detailed solutions, tascCODA is able to provide a trade-off between model sparsity and accuracy, which can be adjusted to reveal credible associations on different levels of the hierarchy. We recapitulated this behavior in synthetic benchmark scenarios, where focusing on low aggregation levels allowed tascCODA to outperform state-of-the-art methods in a differential abundance testing setup, while effects that influenced the majority of features were recovered with greater accuracy when we favored generalizing solutions. The aggregation property further allows for more interpretable models, detecting group-specific changes in the cell lineage or microbial taxonomy. For instance, tascCODA determined B and T cells as the main factors in cell composition changes of the Lamina Propria of Ulcerative Colitis patients, while inflamed epithelial tissue biopsies showed a depletion of Enterocytes.

Second, tascCODA can accommodate any linear combination of normalized covariates, allowing for multi-faceted analysis of complex relationships, while still producing highly sparse and interpretable solutions. On synthetic data, we showed that tascCODA was able to accurately distinguish the influence of two covariates that perturbed the data in different ways. While we did not detect credible relationships with the covariates age, sex and BMI, tascCODA was also able to simultaneously identify characteristic shifts in the gut microbiome for each subtype of Irritable Bowel Syndrome.

The application range of tascCODA extends beyond the taxonomic or expert-derived cell lineage tree structures used in our real data applications. Genetically driven orderings such as phylogenetic trees or cell type hierarchies obtained from clustering algorithms, or approaches aimed at optimizing the predictiveness of the hierarchical grouping (Quinn and Erb, 2019) may provide more accurate results in differential abundance testing (see, e.g., Bichat et al. (2020) for further information).

While tascCODA provides a hierarchically adaptive extension of a classical compositional modeling framework based on a fixed aggregation level, extensions of the method could increase the application range of tascCODA. First, tascCODA does not account for the zero-inflation and overdispersion that is common in microbial abundance data on the OTU/ASV level. We avoided this challenge here by aggregating the amplicon data to the genus level. Accounting for these properties within the model, for example by using a zero-inflated Dirichlet-Multinomial model (Tang and Chen (2019)), the Tweedie family of distributions (Mallick et al. (2021)), or hard thresholding on latent weights (Ren et al. (2020)), would allow for even more fine-grained analyses. Second, the tascCODA model currently places a sparsity-inducing spike-and-slab lasso prior on all included covariates. A natural next step would be to consider some covariates as confounding variables similar to Zhou H. et al. (2021), reducing the number of latent parameters, while restricting results to a few core influence factors. Third, extending known efficient computational methods for inference of spike-and-slab lasso priors (Bai et al. (2020b); Ročková and George (2018)) to be used with our compositional modeling framework could greatly reduce the computational resources required for running tascCODA.

We believe that tascCODA, together with its implementation in Python, represents a valuable addition to the growing toolbox of compositional data modeling tools by providing a unifying statistical way to model and analyze microbial and cell population data in the presence of hierarchical side information.

DATA AVAILABILITY STATEMENT

The model is available as a Python package on github[4]. The datasets used in this study are publicly available on Single Cell Portal (accession ID SCP259) and the Short Read Archive (accession number PRJNA373876). The scripts used for data analysis and benchmark data generation can be found in the tascCODA reproducibility repository[5]. Supplemental data can be downloaded from zenodo[6].

[4]https://github.com/bio-datascience/tascCODA.

[5]https://github.com/bio-datascience/tascCODA_reproducibility.

[6]10.5281/zenodo.5302135.

AUTHOR CONTRIBUTIONS

JO developed tascCODA and conducted the simulation studies and real data analysis. SC processed the 16S rRNA sequencing data and provided biological context. CM supervised the work. JO and CM conceived the statistical model, designed the simulation and out-of-sample prediction studies and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

CM acknowledges core funding from the Institute of Computational Biology, Helmholtz Zentrum München.

ACKNOWLEDGMENTS

We thank Maren Büttner for providing the initial processing steps in the scRNA-seq data analysis. Furthermore, we thank Jennifer S. Labus for kindly sharing additional metadata information on the IBS data. We acknowledge Michael Menden's support in supervising SC during her Master's Thesis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.766405/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv preprint arXiv:1603. 04467.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B (Methodological)* 44, 139–160.

Büttner, M., Ostner, J., Müller, C. L., Theis, F. J., and Schubert, B. (2020). scCODA: A Bayesian Model for Compositional Single-Cell Data Analysis. *Nat. Commun.* 12, 6876.

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020a). Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models. *J. Am. Stat. Assoc.*

Bai, R., Rockova, V., and George, E. I. (2020b). Spike-and-Slab Meets LASSO: A Review of the Spike-And-Slab LASSO. *arXiv [stat.ME]*.

Betancourt, M., and Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In Current Trends in Bayesian Methodology with Applications. Chapman and Hall/CRC, 79–101.

Bichat, A., Plassais, J., Ambroise, C., and Mariadassou, M. (2020). Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control. *Front. Microbiol.* 11, 649.

Bien, J., Yan, X., Simpson, L., and Müller, C. L. (2021). Tree-aggregated Predictive Modeling of Microbiome Data. *Sci. Rep.* 11, 14505. .

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, 581–583.

Chen, J., and Li, H. (2013). Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *Ann. Appl. Stat.* 7.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., et al. (2017). Tensorflow Distributions. *arXiv preprint*. arXiv:1711. 10604

Duan, R., Zhu, S., Wang, B., and Duan, L. (2019). Alterations of Gut Microbiota in Patients with Irritable Bowel Syndrome Based on 16S rRNA-Targeted Sequencing: A Systematic Review. *Clin. Translational Gastroenterol.* 10, e00012.

Duò, A., Robinson, M. D., and Soneson, C. (2018). A Systematic Performance Evaluation of Clustering Methods for Single-Cell Rna-Seq Data. *F1000Res* 7, 1141.

Durbán, A., Abellán, J. J., Jiménez-Hernández, N., Salgado, P., Ponce, M., Ponce, J., et al. (2012). Structural Alterations of Faecal and Mucosa-Associated Bacterial Communities in Irritable Bowel Syndrome. *Environ. Microbiol. Rep.* 4, 242–247.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis. *Microbiome* 2, 15.

Ford, A. C., Lacy, B. E., and Talley, N. J. (2017). Irritable Bowel Syndrome. *N. Engl. J. Med.* 376, 2566–2578.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe* 15, 382–392.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8, 2224.

Gordon-Rodriguez, E., Quinn, T. P., and Cunningham, J. P. (2021). Learning Sparse Log-Ratios for High-Throughput Sequencing Data. *bioRxiv*. doi:10.1101/2021.02.11.430695

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using Single-Cell Genomics to Understand Developmental Processes and Cell Fate Decisions. *Mol. Syst. Biol.* 14, e8046.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2019). A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate. *Brief. Bioinform.* 20, 210–221.

He, S., Wang, L. H., Liu, Y., Li, Y. Q., Chen, H. T., Xu, J. H., et al. (2020). Single-cell Transcriptome Profiling of an Adult Human Cell Atlas of 15 Major Organs. *Genome Biol.* 21, 294.

Holmén, N., Lundgren, A., Lundin, S., Bergin, A.-M., Rudin, A., Sjövall, H., et al. (2006). Functional CD4+CD25high Regulatory T Cells Are Enriched in the Colonic Mucosa of Patients with Active Ulcerative Colitis and Increase with Disease Activity. *Inflamm. Bowel Dis.* 12, 447–456.

Homan, M. D., and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1593–1623.

Human Microbiome Project Consortium (2012). Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* 486, 207–214.

Jeffery, I. B., O'Toole, P. W., Öhman, L., Claesson, M. J., Deane, J., Quigley, E. M. M., et al. (2012). An Irritable Bowel Syndrome Subtype Defined by Species-specific Alterations in Faecal Microbiota. *Gut* 61, 997–1006.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., et al. (2021). A Single–Cell Type Transcriptomics Map of Human Tissues. *Sci. Adv.* 7, 2169.

Kumar, R., Carroll, C., Hartikainen, A., and Martin, O. (2019). ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in python. *Joss* 4, 1143.

Labus, J. S., Hollister, E. B., Jacobs, J., Kirbach, K., Oezguen, N., Gupta, A., et al. (2017). Differences in Gut Microbial Composition Correlate with Regional Brain Volumes in Irritable Bowel Syndrome. *Microbiome* 5, 49.

Lin, H., and Peddada, S. D. (2020). Analysis of Compositions of Microbiomes with Bias Correction. *Nat. Commun.* 11, 3514.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, Functions and Dynamics in the Expanded Human Microbiome Project. *Nature* 550, 61–66.

Luecken, M. D., and Theis, F. J. (2019). Current Best Practices in Single-Cell Rna-Seq Analysis: a Tutorial. *Mol. Syst. Biol.* 15, e8746.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214.

Maier, M. J. (2014). *DirichletReg: Dirichlet Regression for Compositional Data in R*. Research Report Series 125. Vienna, Austria: Vienna University of Economics and Business.

Mallick, H., Chatterjee, S., Chowdhury, S., Chatterjee, S., Rahnavard, A., and Hicks, S. C. (2021). Differential Expression of Single-Cell RNA-Seq Data Using Tweedie Models.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of Composition of Microbiomes: a Novel Method for Studying Microbial Composition. *Microb. Ecol. Health Dis.* 26, 27663.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *Msystems* 3, e00031–18.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J.* 6, 610–618.

McKinney, W. (2010). Data Structures for Statistical Computing in python. In Proceedings of the 9th Python in Science Conference. (Austin, Texas, USA: SciPy).

Nagel, R., Traub, R. J., Allcock, R. J. N., Kwan, M. M. S., and Bielefeldt-Ohmann, H. (2016). Comparison of Faecal Microbiota in Blastocystis-Positive and Blastocystis-Negative Irritable Bowel Syndrome Patients. *Microbiome* 4, 47.

Nesterov, Y. (2009). Primal-dual Subgradient Methods for Convex Problems. *Math. Program* 120, 221–259.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20, 289–290.

Pozuelo, M., Panda, S., Santiago, A., Mendez, S., Accarino, A., Santos, J., et al. (2015). Reduction of Butyrate- and Methane-Producing Microorganisms in Patients with Irritable Bowel Syndrome. *Sci. Rep.* 5, 12693.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* 41, D590–D596.

Quinn, T. P., and Erb, I. (2019). Using Balances to Engineer Features for the Classification of Health Biomarkers: a New Approach to Balance Selection. *bioRxiv*. doi:10.1101/600122

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The Human Cell Atlas. *elife* 6, e27041.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2020). Bayesian Mixed Effects Models for Zero-Inflated Compositions in Microbiome Data Analysis. *Ann. Appl. Stat.* 14, 494–517.

Ročková, V., and George, E. I. (2018). The Spike-And-Slab LASSO. *J. Am. Stat. Assoc.* 113, 431–444.

Round, J. L., and Palm, N. W. (2018). Causal Effects of the Microbiota on Immune-Mediated Diseases. *Sci. Immunol.* 3.

Schliep, K. P. (2010). Phangorn: Phylogenetic Analysis in R. *Bioinformatics* 27, 592–593.

Scott, J. G., and Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Statist.* 38, 2587–2619.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *Plos Biol.* 14, e1002533–14.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., et al. (2013). Single-cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells. *Nature* 498, 236–240.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data. *Elife* 6.

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., et al. (2019). Intra- and Inter-cellular Rewiring of the Human colon during Ulcerative Colitis. *Cell* 178, 714–730.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). Mrna-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6, 377–382.

Tang, Z.-Z., and Chen, G. (2019). Zero-inflated Generalized Dirichlet Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biostatistics* 20, 698–713.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* 9, 5233.

Trapnell, C. (2015). Defining Cell Types and States with Single-Cell Genomics. *Genome Res.* 25, 1491–1498.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G. C. (2019). Accurate Estimation of Cell-type Composition from Gene Expression Data. *Nat. Commun.* 10, 2975–2979.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* 449, 804–810.

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome Data. *BMC Bioinformatics* 18, 94.

Wang, T., and Zhao, H. (2017). A Dirichlet-Tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms. *Biom* 73, 792–801.

Wang, Z., Mao, J., and Ma, L. (2021). Logistic-tree normal Model for Microbiome Compositions. *arXiv [stat.ME]*.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol.* 19, 15.

Yan, X., and Bien, J. (2021). Rare Feature Selection in High Dimensions. *J. Am. Stat. Assoc.* 116, 887–900.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-Species Living Tree Project (LTP)" Taxonomic Frameworks. *Nucl. Acids Res.* 42, D643–D648.

Zhou, C., Zhao, H., and Wang, T. (2021a). Transformation and Differential Abundance Analysis of Microbiome Data Incorporating Phylogeny. *Bioinformatics*.

Zhou, H., Zhang, X., He, K., and Chen, J. (2021b). LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data. *arXiv [stat.ME]*.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 25 February 2022
doi: 10.3389/fgene.2022.777877

ARZIMM: A Novel Analytic Platform for the Inference of Microbial Interactions and Community Stability from Longitudinal Microbiome Study

*Linchen He[1], Chan Wang[2], Jiyuan Hu[2], Zhan Gao[3], Emilia Falcone[4], Steven M. Holland[4], Martin J. Blaser[3] and Huilin Li[2]**

[1]Novartis Pharmaceuticals Corporation, East Hanover, NJ, United States, [2]Division of Biostatistics, Department of Population Health, New York University School of Medicine, East Hanover, NY, United States, [3]Center for Advanced Biotechnology and Medicine, Rutgers University, New Brunswick, NJ, United States, [4]Division of Intramural Research, Immunopathogenesis Section, NIAID, NIH, Bethesda, MD, United States

OPEN ACCESS

Edited by:
Himel Mallick, Merck, United States

Reviewed by:
Boyu Ren, Dana–Farber Cancer Institute, United States
Siyuan Ma, University of Pennsylvania, United States
Qiwei Li, The University of Texas at Dallas, United States

***Correspondence:**
Huilin Li
huilin.li@nyulangone.org

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *16 September 2021*
Accepted: *31 January 2022*
Published: *25 February 2022*

Citation:
He L, Wang C, Hu J, Gao Z, Falcone E, Holland SM, Blaser MJ and Li H (2022) ARZIMM: A Novel Analytic Platform for the Inference of Microbial Interactions and Community Stability from Longitudinal Microbiome Study. Front. Genet. 13:777877. doi: 10.3389/fgene.2022.777877

Dynamic changes of microbiome communities may play important roles in human health and diseases. The recent rise in longitudinal microbiome studies calls for statistical methods that can model the temporal dynamic patterns and simultaneously quantify the microbial interactions and community stability. Here, we propose a novel autoregressive zero-inflated mixed-effects model (ARZIMM) to capture the sparse microbial interactions and estimate the community stability. ARZIMM employs a zero-inflated Poisson autoregressive model to model the excessive zero abundances and the non-zero abundances separately, a random effect to investigate the underlining dynamic pattern shared within the group, and a Lasso-type penalty to capture and estimate the sparse microbial interactions. Based on the estimated microbial interaction matrix, we further derive the estimate of community stability, and identify the core dynamic patterns through network inference. Through extensive simulation studies and real data analyses we evaluate ARZIMM in comparison with the other methods.

Keywords: autoregressive, longitudinal microbiome data, microbial community stability, mixed-effects model, zero-inflated, network analysis, microbial interaction, absolute abundance

INTRODUCTION

The human microbiota, a diverse array of microbial organisms living in and on human bodies, form a dynamic ecosystem that plays a critical role in human health. While temporally stable microbial communities are observed among healthy adults (Faith et al., 2013), the fluctuation of microbiome has been linked to increasing frailty (Jackson et al., 2016) and declining immune function of hosts (Claesson et al., 2012), and diseases such as inflammatory bowel disease (Martinez et al., 2008; Zuo and Ng, 2018), colorectal cancer (Scanlan et al., 2008; Uronis et al., 2009), and irritable bowel syndrome (Maukonen et al., 2006; Carroll et al., 2012). When a microbial community changes in response to an external perturbation, it undergoes a dynamic process and tends to evolve toward another stable state (**Figure 1**). This dynamic process is stochastic and varies according to the type and strength of perturbation, the community stability prior to the perturbation, and other subject-level relevant features. The recent rise in longitudinal studies, in which microbial samples are collected repeatedly over time, offers unique insights into the responses of such communities to perturbations and the associated dynamic patterns. For example, in our ongoing microbiome study

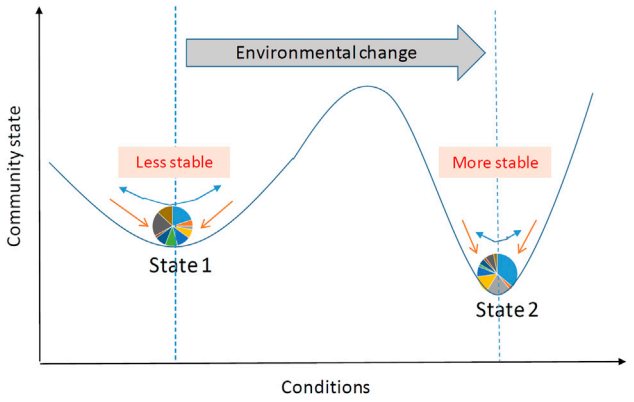


FIGURE 1 | Schematic of the evolution of microbial community states in response to external perturbation. External perturbation (blue arrows) can affect microbial community composition (shown in a pie chart), defined as a community state. For each state, the ball-in-basin diagram portrays stability measured by the variance in the stationary distribution of the location of the ball. White arrows indicate the reaction of microbial community to the perturbation.

evaluating the effects of antibiotic exposure as a short-term perturbation on microbial, immune, and metabolic physiology (MIME study), we are interested in determining how differently the microbial community responds to the antibiotic treatment.

Human microbiota studies have been accelerated by the advent of next-generation sequencing technologies which enabled the quantification of the composition of microbiomes, often by two common sequencing approaches—16S rRNA marker gene sequencing and shotgun metagenomics sequencing (Woo et al., 2008). There are pros and cons to each of those techniques, which are discussed in recent reviews (Shankar, 2017; Gilbert et al., 2018). But for both methods, because of the varying sequencing read counts obtained across samples, it is necessary to employ various normalization tools to convert raw counts data into relative abundances (Knight et al., 2018). However, the dependency of the compositional components greatly hampers the interpretation of microbiota changes in longitudinal studies. There is reason to believe that the absolute abundances of bacteria are biologically meaningful measures, especially in the study of microbial interactions. Thus, in our MIME study, we use an independent quantitative polymerase chain reaction (qPCR) technology (Nadkarni et al., 2002; Ott et al., 2004; Kim et al., 2013) to quantify total bacterial load per unit sample, and then use these data to estimate absolute bacterial abundance by combining them with the relative abundance values obtained from 16S rRNA or shotgun sequencing methods. This MIME study motivated us to develop analytical methods to investigate microbial interaction and community stability after a strong external perturbation, and identify core active microbial taxa by modeling the absolute abundances of bacteria.

Although many well-developed statistical tools are widely used for assessing the diversity of microbial communities and its composition, there are only a few methods available for inferring the ecological networks of microbial communities. Here we briefly review the well-developed statistical methods for studying the dynamic microbial systems and their limitations.

A Bayesian network contains a set of multivariate joint distributions that exhibit certain conditional independences and a directed and acyclic graph that encodes conditional independences among random variables. If the dependence relationships repeat and the signals at a certain time point only depend on the signals from previous time points, then the whole network can be formulated as a dynamic Bayesian network (DBN) (Russell and Norvig, 2002) representation. McGeachie et al. (McGeachie, 2016) constructed a simplified two-stage DBN which uses a Markov assumption that the observed values at time $t+1$ are independent of those at earlier time points ($t-1$ and earlier) given the variable values at time t. Lugo-Martinez et al. presented a computational pipeline which first aligns the data collected from all individuals, and then learns a dynamic Bayesian network from the aligned profiles (Lugo-Martinez et al., 2019). However, DBN has several limitations in analyzing the longitudinal microbial data. 1) It can only model the microbial community subject-by-subject. 2) DBN cannot handle the excess zeros in microbiome data. Most methods remove the taxa whose relative abundances exhibit zero entry (i.e., not present in a measurable amount at one or more of the measured time points) before the downstream analysis. 3) The assumed distributions are unrealistic. E.g. all continue variables are assumed to be normally distributed. 4) The computational cost is relatively high, since parent nodes are added sequentially for each bacterial node. Additionally, the maximum number of possible parents is imposed, which is not realistic. 5) Due to sampling and sequencing limitations, the compositionality bias in microbiome data may also cause inaccurate estimation of parameters. The existing methods ignore this compositionality bias, making parameter estimates difficult to interpret. 6) Irregular sampling time may also result in inaccurate parameter estimation. Therefore, it is advised to cautiously interpret the findings from DBN (Faust and Raes, 2012; Gerber, 2014).

The classical Lotka-Volterra equation has been used to model simple system such as two species in a predator-prey relationship, where the interactions are strictly assumed to be competitive. The generalized Lotka-Volterra (gLV) equations extend the classical predator-prey (Lotka-Volterra) equations, where the interacting species might have a wide range of relationships including competition, cooperation, or neutralism. Assuming that the interaction (or the effect) of one species with another can be modeled by the corresponding coefficient in the equation, gLV equations provide a framework to analyze and simulate microbial populations. Mounier et al. used the gLV equations to model the interaction between bacteria and yeast in a cheese microbiome (Mounier et al., 2009). Other microbiome studies further extended and implemented the gLV equations (Marino et al., 2014; Dam et al., 2016; de Vos et al., 2017; Venturelli et al., 2018).

Many software are available for applying gLV modeling on microbial time series data, such as LIMITS (Fisher and Mehta, 2014), MetaMis (Shaw et al., 2016), and MDSINE (Bucci et al., 2016a). LIMITS and MetaMis can be implemented to construct microbial interactions using the longitudinal microbiome data

from one subject. MDSINE can jointly analyze multiple time series, but requires Matlab programming. Web-gLV (http://web.rniapps.net/webglv) can be used for modeling, visualization, and analysis of microbial populations, but can only handle limited number of samples. In summary, there are several limitations of gLV in analyzing the longitudinal microbial data. 1) gLV based models capture the interactions using a single averaged effect, thus they are not well-suited for noisy data. 2) Some methods estimate almost all possible edges without incorporating variable selection techniques. 3) gLV estimates the growth rate of each taxon marginally, therefore, ignores the intrinsic dynamic correlations of the repeated measurements. 4) gLV does not account for random processes which form essential part of any biological system. 5) With the increased number of species and time span of prediction, the simulation output is prone to numerical errors. For example, Web-gLV can only simulate a maximum of 10 species at a time for at most 100 time points. 6) As DBN, gLV is not suitable for sparse, compositional, and irregular sampled microbiome data.

In Ives et al. (Ives et al., 2003), the stability of a microbial community is determined by three key interrelated components of microbial community structure: diversity, species composition, and interaction pattern among species. They viewed the dynamics of a microbial community as a stochastic process and proposed to use a first-order multivariate autoregressive process [MAR (1)] time-series model to disentangle the effects of these three components on community stability and to estimate the stability properties of a community by estimating the strengths of interactions between species. This method is widely used to estimate the stability of ecosystems (e.g., lake, ocean) based on culture-dependent microbial data (Carpenter et al., 2011; Shade et al., 2013). Usually a few (four or five) key microbes are detected with high frequency in each ecosystem in time-series measurements over a long period, and their abundances are rarely zero. In contrast, our MIME study will yield microbiome data from approximately nine time points over half a year from 80 subjects in three groups in the complete study—a relatively smaller number of repeated microbiome samples but from a relatively larger number of microbial communities (subjects) than what would be the case for an ecosystem study. Moreover, the 16S rRNA sequencing and qPCR methods used in this study provide absolute abundances for a staggering number of taxa, which include a large number of zero values. Because the MAR modeling methods require the normality assumption, they are not appropriate for analyzing data from sequence-based longitudinal microbiome studies. Therefore, we propose an autoregressive zero-inflated mixed effects model (ARZIMM) to address the special features of data instead. Its novelties are threefold. First, we propose to use a zero-inflated Poisson autoregressive model to model the excessive zero abundances and the non-zero abundances separately. Second, the random effects in the proposed model can investigate the underlining dynamic pattern shared within the group. Third, the employment of regularization techniques and network inference in our model enables the identification of the core dynamic patterns. The proposed ARZIMM estimates the strength of interactions between taxa, which is required to estimate the stability properties of a community, and identify key active taxa efficiently by using all of the longitudinal sequencing data. ARZIMM has been implemented in an open-source software package (https://github.com/Hlch1992/ARZIMM), and provides a useful tool for formulating, understanding, and implementing longitudinal microbiome data analysis.

In the following Material and Method section, we introduce the ARZIMM framework, discuss the quantification of microbial stability based on the estimated microbial interaction matrix, and investigate the conditions under which there exist a strict-sense stationary distribution. Then in the Result section, we evaluate ARZIMM using extensive simulation studies to show that it outperforms the conventional methods, and apply ARZIMM to the MIME study to illustrate network visualization and inference. In the end, we conclude with Discussion section.

MATERIALS AND METHODS

ARIZMM Model

As illustrated in **Figure 2**, ARZIMM can be considered as a two-part model which comprises a logistic component and an autoregressive component. To address zero inflation, we consider the zero-inflated mixture model because it assumes both sampling zeros (due to the low sequencing depth) and structural zeros (being truly absent) exist in the data. Specifically, the logistic component models the structure zeros of taxa in the samples, and the autoregressive component models the non-structure-zero abundances of the taxa under the assumption that the changes in abundances from time $t-1$ to time t depend only on the observed abundances at time $t-1$ and other time-independent covariates, and the observed abundances before time $t-1$ have no direct effect. Since the goal of ARZIMM is to characterize microbial interactions and community stability during a short period after a strong external perturbation like the antibiotic usage in our MIME study, we assume there are no other time-dependent factors exist to affect the microbial stability.

Notation and Model Specifications

Let Y_{imt} denote the observed absolute abundance of bacterial taxon m $(m = 1, ..., M)$ for subject i at time t $(i = 1, 2, ..., n,\ t = 1, ..., T_i)$, and we model Y_{imt} with a conditional mixture distribution as follow:

$$Y_{imt}\Big|\nu_{i(t-1)} \sim \begin{cases} 0 & p_{im} \\ F\left(y_{imt}\middle|\nu_{i(t-1)};\theta_{itm},\ \phi_m\right) & 1-p_{im} \end{cases} \quad (1)$$

where $\nu_{i(t-1)}$ represents all information that is known at time $(t-1)$ for individual i, including the observed absolute abundance $Y_{im(t-1)}$ and later defined coviariates $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$. The parameter p_{im} represents the probability of the observation Y_{imt} being structural zero and is assumed time independent. Furthermore, F is assumed to be an exponential dispersion family distribution with the canonical parameter θ_{imt} and the dispersion parameter ϕ_m. Both Poisson and negative binomial (NB) distributions can be used as to model absolute abundance. Below we illustrate the detailed modelling using Poisson model.

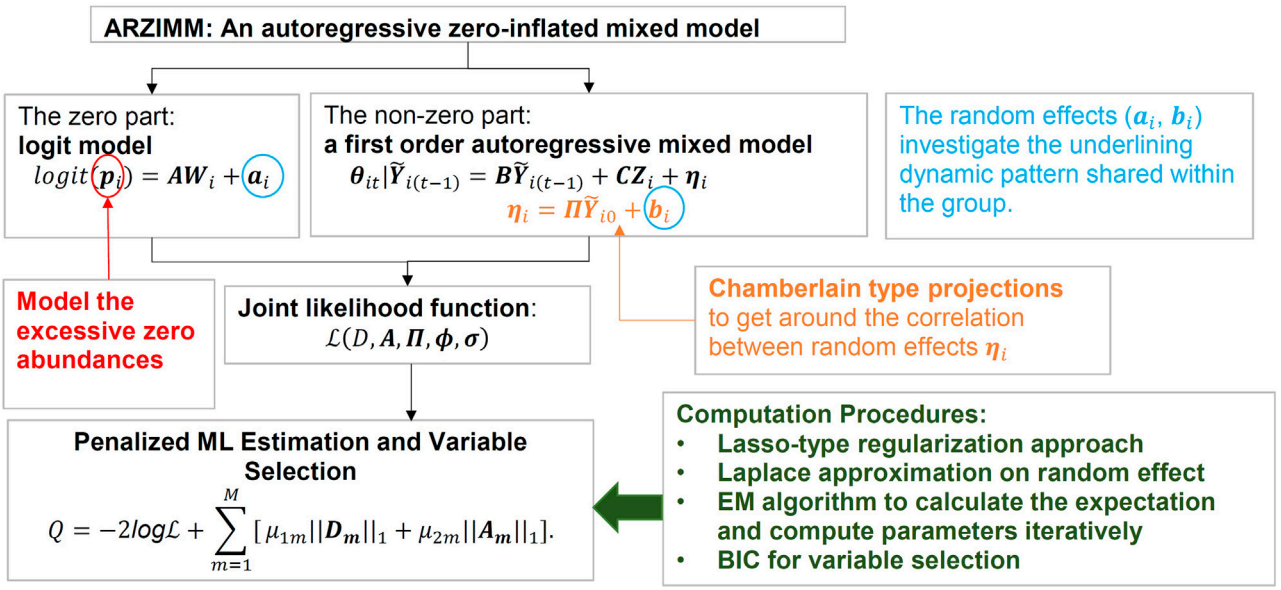


FIGURE 2 | Graphical representation of ARZIMM model and analytic techniques.

The mixture probability parameters $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iM})'$ are modeled by the logistic regression:

$$log\,it\,(\boldsymbol{p}_i) = \boldsymbol{A}\boldsymbol{W}_i + \boldsymbol{a}_i \tag{2}$$

where $\boldsymbol{W}_i = (1, w_{i1}, \ldots, w_{il})'$ consists of intercept and l time independent covariates for individual i, the parameter $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M)'$ is an $M \times (l+1)$ matrix whose elements A_{mj} is the effect of covariate j on the zero proportion of taxon m. $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iM})'$ is an $M \times 1$ vector of random intercepts to model the within-subject heterogeneity of being zero for individual i and has the joint multivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\mathbf{a})$.

The canonical parameters for Poisson distribution is $\theta_{imt} = \log E(Y_{imt})$. We introduce the auto-regressive model by relating $\boldsymbol{\theta}_{it} = (\theta_{i1t}, \ldots, \theta_{iMt})'$ to the i^{th} individual's observed log-transformed absolute abundance vector at time $t-1$: $\tilde{\boldsymbol{Y}}_{i(t-1)} = (\log(Y_{i1(t-1)}+1), \ldots, \log(Y_{iM(t-1)}+1))'$ (where the pseudo count 1 is added to avoid the undefined logarithm when the absolute abudance is zero), and $\boldsymbol{Z}_i = (1, Z_{i1}, \ldots, Z_{iq})'$, the intercept and q time-independent covariates of individual i by

$$\boldsymbol{\theta}_{it}\big|\tilde{\boldsymbol{Y}}_{i(t-1)} = \boldsymbol{B}\tilde{\boldsymbol{Y}}_{i(t-1)} + \boldsymbol{C}\boldsymbol{Z}_i + \boldsymbol{\eta}_i \tag{3}$$

where $\boldsymbol{B}$ is an $M \times M$ matrix whose element B_{mj} gives the effect of the abundance of taxon j on the growth rate of taxon m, $\boldsymbol{C}$ is an $M \times (q+1)$ matrix whose element C_{mj} gives the effect of covariate j on taxon m, and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iM})'$ is time-independent random intercepts. Note that, as an autoregressive model, $\boldsymbol{\eta}_i$ is correlated with the fixed effect $\tilde{\boldsymbol{Y}}_{i(t-1)}$ and this dependency can be tracked all the way back to the initial observation $\tilde{\boldsymbol{Y}}_{i0}$. Because the standard random effects model has assumption that the random effects are independent to the other covariates in the model, in order to derive the random effect type maximum likelihood (ML) estimators, we use the Chamberlain type projections (Chamberlain, 1982) to get around this correlation. Specifically, we project $\boldsymbol{\eta}_i$ onto the time 0 observations $\tilde{\boldsymbol{Y}}_{i0}$ by:

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}\tilde{\boldsymbol{Y}}_{i0} + \boldsymbol{b}_i \tag{4}$$

where $\boldsymbol{\Pi}$ is an $M \times M$ matrix with $\mathrm{diag}(\boldsymbol{\Pi}) = (\pi_1, \ldots, \pi_M)'$ and off-diagonal components being zero. The components of $\boldsymbol{\Pi}$ represent how much variation in η_i is due to the dependence on subject i's initial value $\tilde{\boldsymbol{Y}}_{i0}$. $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iM})'$ is an $M \times 1$ vector, representing the independent subject-specific random effect and follows a joint multivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\mathbf{b})$.

In the model, our primary interest is to estimate matrix $\boldsymbol{B}$, which measures the strengths of interactions between taxa. For a microbial community with a given number of species, its stability or dynamics status depends on the changes in the species' population growth rates due to perturbation, which immediately cause the changes in the population growth rates of other species via species-species interactions (Ives et al., 2000). Interaction between species can be viewed as a filter that amplifies the variability in species' population growth rates caused by perturbation.

Note that we choose Poisson distribution because of its nice stationary distribution property in the autoregressive model which is crucial for our following stability investigation. To deal with the over-dispersion of microbiome data, we implemented the quasi-Poisson model (Ver Hoef and Boveng, 2007) in the simulation and real data analysis.

Penalized ML Estimation and Variable Selection

To define the joint likelihood of the longitudinal microbial absolute abundance data $\mathbf{Y}_{it}$, we assume that the vector of

time independent random effects $c_i = (a_i', b_i')'$ underlies both the zero and autoregressive generative processes and these random effects account for the within-subject group heterogeneity in the multivariate logistic component and the multivariate autoregressive component. Denote $\boldsymbol{D} = (\boldsymbol{B}, \boldsymbol{C}) = (\boldsymbol{D}_1, ..., \boldsymbol{D}_M)'$, $\phi = (\phi_1, ..., \phi_M)'$, and $1_{[\cdot]}$ as the indication function that when $[\cdot]$ meets, $1_{[\cdot]} = 1$, otherwise, $1_{[\cdot]} = 0$. Formally, we have the joint likelihood function as:

$$\mathcal{L}(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\Pi}, \boldsymbol{\phi}, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \int \left\{ \left[\prod_{t=1}^{t_i} \prod_{m=1}^{M} f_y\left(y_{itm} | \theta_{itm}(b_{im}), \phi_m, p_{im}(a_{im})\right) \right] g(\boldsymbol{c}_i | \boldsymbol{\Sigma}(\boldsymbol{\sigma})) \right\} dc_i \tag{5}$$

where f_y is the conditional probability density function and given as

$$f_y\left(y_{itm} | \theta_{itm}(b_{im}), \phi_m, p_{im}(a_{im})\right) = \left[p_{im} + (1 - p_{im}) f\left(y_{itm} = 0 | \theta_{itm}, \phi_m\right)\right]^{1_{[y_{itm}=0]}} \times \left[(1 - p_{im}) f\left(y_{itm} | \theta_{itm}, \phi_m\right)\right]^{1_{[y_{itm} \neq 0]}} \tag{6}$$

The function $g(\boldsymbol{c}_i | \boldsymbol{\Sigma}(\boldsymbol{\sigma}))$ is the joint distribution of $\mathbf{c_i}$, and $\boldsymbol{\Sigma}(\boldsymbol{\sigma}) = \begin{bmatrix} \boldsymbol{\Sigma_a} & \boldsymbol{\Sigma_{ab}} \\ \boldsymbol{\Sigma_{ab}} & \boldsymbol{\Sigma_b} \end{bmatrix}$ represents the corresponding $2M \times 2M$ covariance matrix, where $\boldsymbol{\sigma}$ accounts for all unique non-zero elements of Σ. For the model and computational simplicity, we assume $Cov(\boldsymbol{a}_i, \boldsymbol{b}_i) = \Sigma_{ab} = 0$, i.e. $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are independent.

Assuming that the true underlying fixed effects $\boldsymbol{A}$ and $\boldsymbol{D}$ are sparse, we advocate a Lasso-type approach, which adds an ℓ_1 -penalty for the fixed-effects to the likelihood function. Thus, we consider the following objective function:

$$Q = -2 \log \mathcal{L} + \sum_{m=1}^{M} \left[\mu_{1m} ||\boldsymbol{D}_m||_1 + \mu_{2m} ||\boldsymbol{A}_m||_1\right]. \tag{7}$$

Maximization of the penalized log-likelihood function corresponding to **Eq. 7** with respect to $(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\Pi}, \boldsymbol{\phi}, \boldsymbol{\sigma})$ is a computationally challenging task. This is mainly because both integrals with respect to the random effects and the zero-inflated structure do not have analytical solutions. Following the conventional methods, we propose to implement a Laplace approximation on the integral of random effects in **Eq. 7** and use the Expectation-Maximization (EM) algorithm to calculate the expectation and compute parameters iteratively, in which the label of zero is treated as "missing data". The tuning parameters are selected using Bayesian information criterion (BIC).

Stability Properties

The existence of a stationary distribution has been investigated for the log-linear Poisson auto-regression model based on the perturbation technique (Fokianos and Tjøstheim, 2011). Here, we prove the existence of a stationary distribution of a zero-inflated Poisson mixed-effect auto-regression model in Theorem 1 utilizing the theory of Markov chains which has been proposed to prove the existence of a stationary distribution of a general class of time series count models (Douc et al., 2013). The detailed proof is provided in the **Supplementary Material Section S3**.

Theorem 1. Assuming that time-independent parameters $\boldsymbol{\eta}_i$ and p_i are known, if all eigenvalues of matrix $\boldsymbol{B}$ lie inside the unit circle, a strict-sense stationary ergodic process $\{\boldsymbol{Y}_{it}\}_{t \in N}$ will exist, where $\boldsymbol{N}$ denotes the set of natural numbers.

With this Theorem, we can first show that for a microbial community, its dynamic process $\{\boldsymbol{Y}_{it}\}_{t \in \boldsymbol{N}}$ has a stationary distribution by proving that all eigenvalues of matrix $\boldsymbol{B}$ lie inside the unit circle. Then, following Ives et al. (Ives et al., 2003), we consider the return rate and reactivity as two stability measures based on the variability of the stationary distribution for MAR (1) model. Specifically, return rate depends on the rate at which the perturbed microbial community approaches the stationary distribution and reactivity, and assesses how strongly population-level microbiome abundances are pulled towards the mean of the stationary distribution. Both are bounded by the largest eigenvalue of B, denoted by $\max(\boldsymbol{\lambda}_B)$. In general, a smaller $\max(\boldsymbol{\lambda}_B)$ indicates the perturbed microbial community approaches its stationary distribution faster, or a system is less reactive, then the microbial community is more stable. The detailed proof is deferred in the **Supplementary Material Section S3.2**.

Based on the theory in Ives et al. (Ives et al., 2003), for a community with multiple species, the covariance matrix of the stationary distribution depends on the covariance matrix of the process error and the interactions between species captured in the matrix $\boldsymbol{B}$. As illustrated in **Figure 1**, when the external perturbation(blue arrow) acts on the community, the ball(microbial community) sitting in a deep bowl in state 2 which represents a relatively stable system, will return to its stationary state faster than the ball sitting in a shallow bowl in state 1 which represents a less stable system. In a stable system, the variance of stationary distribution is only slightly greater than the variance of process error and the variance of species interaction is small. In contrast, in a less stable system, the species interaction will amplify the environmental variance and create large variance in the stationary distribution, therefore the variance of species interaction is large, assuming the process errors are similar in the compared two states. Thus, the difference between the variances of stationary distribution of different communities can be attributed to species interactions. The smaller of the variance of matrix B, the more stable of the study microbial community.

RESULTS

Simulation Study

We have conducted extensive simulation studies to evaluate the performance of ARZIMM in both model fitting and variable selection by comparing it with the competing methods: penalized Poisson auto-regression (Poisson), penalized log-normal multivariate auto-regression (MAR), and extended generalized Lotka-Volterra (gLV) equations using Bayesian algorithm (MDSINE) (Bucci et al., 2016b). The brief descriptions of these methods are provided in the **Supplementary Material Section S2**.

Simulation Design

We generated the longitudinal absolute abundances from zero-inflated Poisson distribution with parameters p_{im} and θ_{imt} for

each taxon. Since our focus is on the estimation of the interaction matrix $\boldsymbol{B}$, which depends on the non-zero part, we adopted a simple simulation design for the zero inflation proportions $\boldsymbol{p}_i = (p_{i1}, ..., p_{iM})'$. We ignored the individual variations in $\boldsymbol{p}_i$ by dropping the random effect term $\boldsymbol{a}_i$ in **Eq. 2**. With model $logit(\boldsymbol{p}_i) = \boldsymbol{A}\boldsymbol{W}_i$ and by controlling the values of $\boldsymbol{W}$ and $\boldsymbol{A}$ respectively, we set the zero inflation proportions $\boldsymbol{p}_i$ for 20 taxa to mimic the observed sparsity in real data as

$$\boldsymbol{p}_i = (0.72,\ 1.00,\ 0.96,\ 0.34,\ 0.50,\ 0.56,\ 0.94,\ 0.84,\ 0.98,\ 1.00,\ 0.78,\ 0.68,\ 0.96,\ 1.00,\ 0.38,\ 0.56,\ 0.82,\ 1.00,\ 0.28,\ 1.00)' \quad (8)$$

The detailed values of $\boldsymbol{W}$ and $\boldsymbol{A}$ are provided in the **Supplementary Material Section S4**. We generated the non-zero absolute abundances from Poisson distribution with their $\boldsymbol{\theta}_{it} = (\theta_{i1t}, ..., \theta_{iMt})'$ defined as $\boldsymbol{\theta}_{it} = \boldsymbol{B}\tilde{\boldsymbol{Y}}_{i(t-1)} + \boldsymbol{b}_0 + \boldsymbol{b}_i$, where the intercept $\boldsymbol{b}_0$ was set to be the mean log-transformed non-zero absolute abundances of taxa in MIME real data, and the random effects $\boldsymbol{b}_i \sim \mathcal{N}(0, diag(\boldsymbol{\Sigma}_b))$ with $diag(\boldsymbol{\Sigma}_b) \sim 10^{\mathcal{N}(-1.5,0.5)}$. We assumed that the interaction matrix $\boldsymbol{B}$ was sparse by randomly selecting 5% of its elements to be non-zero. Three interaction matrices were considered with varied informative absolute effect strengths: high ($B^H_{jm} \sim 10^{\mathcal{N}(-0.5.0.5)}$), medium ($B^M_{jm} = \sqrt{0.1}\beta^H_{jm}$), and low ($B^L_{jm} = 0.1B^H_{jm}$), for the non-zero elements B_{jm}. In addition, we designed four simulation scenarios: Scenario 1 with $diag(\boldsymbol{\Sigma}_b) = 0$ and $\boldsymbol{p}_i = 0$, considered as the benchmark situation where subjects are homogeneous and taxa are all presented; Scenario 2 with $diag(\boldsymbol{\Sigma}_b) \sim 10^{\mathcal{N}(-1.5.0.5)}$ and $\boldsymbol{p}_i = 0$, where subjects are heterogeneous and taxa are all presented; Scenario 3 with $diag(\boldsymbol{\Sigma}_b) = 0$ and $\boldsymbol{p}_i$ as in **(8)**, where subjects are homogeneous and taxa have zero inflated structure; and Scenario 4 with $diag(\boldsymbol{\Sigma}_b) \sim 10^{\mathcal{N}(-1.5.0.5)}$ and $\boldsymbol{p}_i$ as in **(8)**, where subjects are heterogeneous and taxa have zero inflated structure.

In each scenario, we generated 500 independent repetitions for $n = 20$ or 50 subjects, $T = 10$ or 20 time points, and $M = 20$ taxa for each sample to evaluate the performance of ARZIMM.

Simulation Results

We first compared the model fittings of ARZIMM, Poisson, and MAR methods using mean normalized squared error score (MNSES), as suggested in the prior studies (Carroll and Cressie, 1997; Liesenfeld et al., 2006; Czado et al., 2009; Tkacz et al., 2018a). MNSES is defined as $\frac{1}{n \times T \times M}(\frac{y_{imt} - \hat{y}_{imt}}{\hat{\sigma}_{y_{imt}}})^2$ with $\hat{y}_{imt}$ being the estimated y_{imt} and $\hat{\sigma}_{y_{imt}}$ being the estimated standard error of y_{imt}. The closer the MNSES is to 1, the better model fitting the method has. Since MDSINE only provides the estimates of interactions among species without their variance estimates, it was excluded from this comparison. **Table 1** and **Supplementary Table S1** summarize the median and interquartile range (IQR) of MNSES over 500 replications for these three methods. Overall, the medians of MNSES for ARZIMM are all around the expected value of 1 in various settings across four scenarios, which indicates the good fitness and robustness of ARZIMM in dealing with excess zeros and the correlation among repeated measures at the same time, as well as its satisfying estimation accuracy on the microbial interaction parameters. However, the other two methods: Poisson and MAR, both exhibit inferior performance. The Poisson model is only competent in Scenario 1, when subjects are homogeneous and no excess zeros are present. In Scenarios 2-4, when any factor, excess zero or subject heterogeneity, presents, the predicted values based on the Poisson model deviate greatly from the observed values. Comparing the considered two factors, Poisson model is more sensitive to the subject heterogeneity and presents larger deviations with it. Due to the invalid normality assumption and lack of consideration of the correlation among the longitudinal measurements, the MAR model exhibits the worst performance among three methods with enormous deviation especially in Scenarios 3 and 4, which confirms the inappropriateness of using conventional statistical methods which require the normality assumption to analyze the microbiome data.

Next, we evaluated the variable selection performance for ARZIMM, Poisson, MAR, and MDSINE in terms of true positive rate (TPR; mathematically equals to the power) and false positive rate (FPR; mathematically equals to the type I error). Specifically, TPR quantifies the probability of a significant interaction identified by one method given that the interaction effect is truly nonzero; and FPR quantifies the probability of a significant interaction identified by one method given that the interaction effect is truly zero. The simulation results for 50 subjects with 20 time points are summarized in **Figure 3** and all the other simulation results with different subject numbers and time points are deferred to **Supplementary Figure S1**, because they have a similar pattern as seen in **Figure 3**. **Figure 3** shows that the FPRs of ARZIMM are all at or below the nominal level (5%) across different simulation regimes and effect sizes, and its TPR estimates exhibit a sensible and consistent pattern as they increase as the interaction effect gets stronger across four scenarios. As expected, the FPR and TRP estimates of Poisson and ARZIMM models are coincident under Scenario 1, because when subjects are homogeneous and taxa don't have excess zeros, ARZIMM model is reduced to Poisson model. However, in Scenarios 2-4, because simple Poisson model fails to take care of the excess zeros or subject heterogeneity, it suffers from the inflated false positives, while ARZIMM does not. For the other two methods, both MAR and MDSINE perform poorly on controlling false positive rates for all simulation scenarios, because MAR fails to fit the skewed and highly sparse microbiome data, while MDSINE captures the interactions based on the averaged effect over subjects in a group but completely ignores the randomness at the subject level process which is the essential characteristic of any biological system. In summary, ARZIMM outperforms the other competitors in handling the excess zeros and subject heterogeneity well with controlled FPR and satisfactory TPR.

To further investigate the performance of informative interaction selection, we calculate Matthew correlation coefficient (MCC), defined as $\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, and F-score, defined as $\frac{TP}{TP+(FP+FN)/2}$, where TP gives the number of selected interactions being true positive, FP gives the number of selected interactions being false positive, TN gives the number of unselected interactions being true

TABLE 1 | Simulation results for all settings under scenario 1 and 4. Poisson refers to the penalized Poisson autoregression model and MAR refers to penalized log-normal multivariate autoregression model. The reported value is median (IQR) of mean normalized squared error score (MNSES) calculated over 500 simulations for each setting. *n* refers to the number of subjects, and *T* refers to the number of time points. Scenarios 2 and 3 are deferred to Supplementary Material.

Methods		**ARZIMM**			**Poisson**			**MAR**		
Effect size		**High**	**Median**	**Low**	**High**	**Median**	**Low**	**High**	**Median**	**Low**
Scenario 1										
n	T	Median (IQR)								
20	10	0.98 (0.97–0.99)	0.98 (0.97–0.99)	0.98 (0.97–0.99)	1.00 (0.99–1.01)	0.99 (0.984–1.00)	1.00 (0.99–1.00)	47 (33–77)	50 (35–80)	52 (37–80)
50	20	0.99 (0.99–1.00)	0.99 (0.99–1.00)	0.99 (0.99–1.00)	1.00 (1.00–1.01)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	123 (86–192)	115 (78–187)	114 (80–177)
Scenario 4										
n	T	Median (IQR)								
20	10	0.95 (0.87–2.30)	0.92 (0.86–1.10)	0.91 (0.86–1.02)	30.59 (21.90–41.51)	18.87 (15.26–21.95)	18.46 (14.77–21.80)	30071.82 (8984–133153)	29390 (13929–77371)	22435 (10251–50171)
50	20	1.09 (1.05–1.20)	0.92 (0.90–0.93)	0.85 (0.85–0.86)	40.31 (36.80–43.63)	31.08 (30.25–31.76)	30.30 (29.65–30.95)	211141 (118860–473551)	110656 (80809–202068)	93579 (66942–167227)

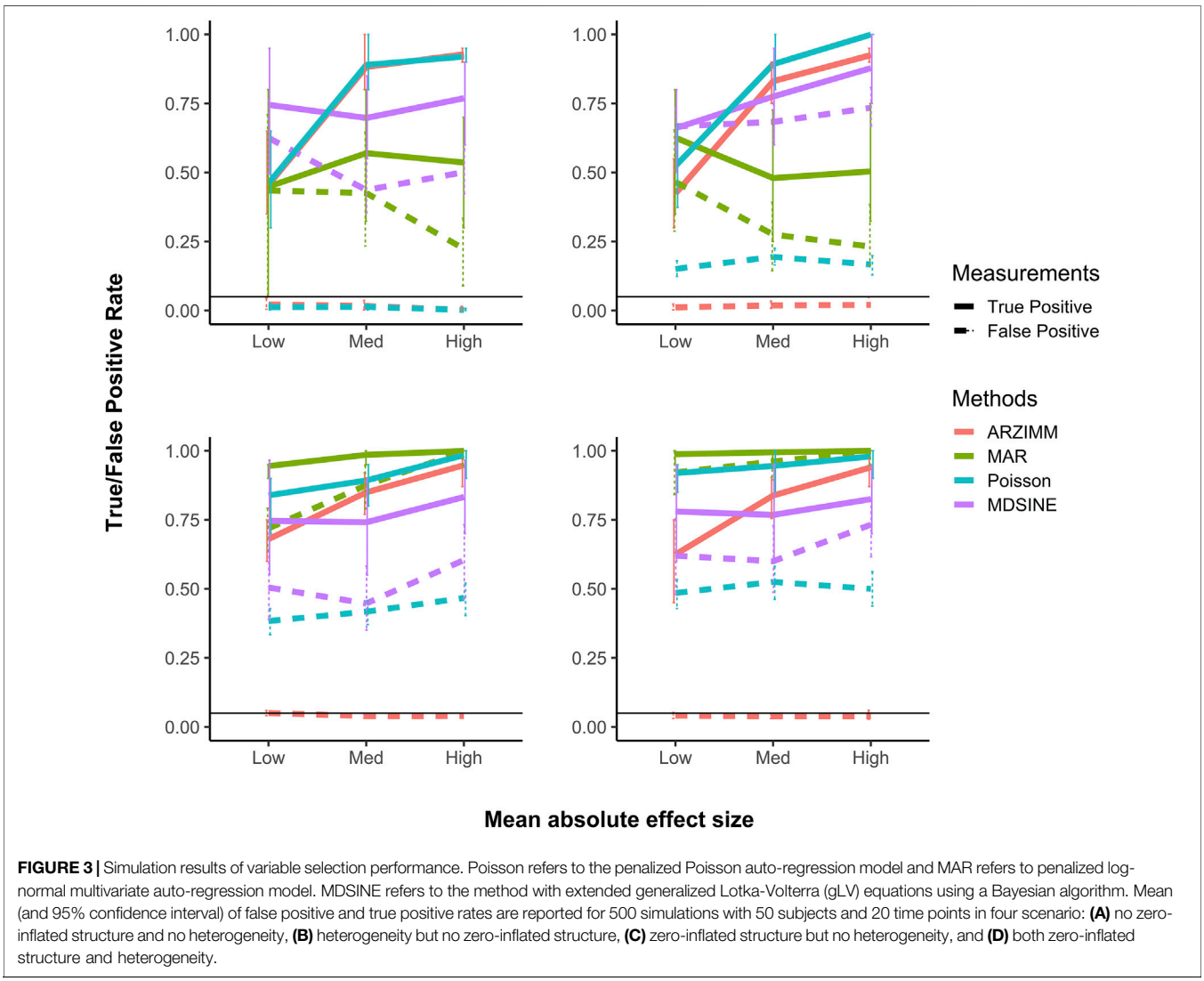


FIGURE 3 | Simulation results of variable selection performance. Poisson refers to the penalized Poisson auto-regression model and MAR refers to penalized log-normal multivariate auto-regression model. MDSINE refers to the method with extended generalized Lotka-Volterra (gLV) equations using a Bayesian algorithm. Mean (and 95% confidence interval) of false positive and true positive rates are reported for 500 simulations with 50 subjects and 20 time points in four scenario: **(A)** no zero-inflated structure and no heterogeneity, **(B)** heterogeneity but no zero-inflated structure, **(C)** zero-inflated structure but no heterogeneity, and **(D)** both zero-inflated structure and heterogeneity.

negative, and FN gives the number of selected interactions being false negative. MCC ranges from −1 to 1, where value 1 indicates perfect agreement between truth and selection, value −1 indicates perfect disagreement, and value 0 indicates that the selection is random with respect to the truth. F-score ranges from 0 to 1, where value 1 indicates that there are neither false negatives nor false positives and value 0 only indicates no true positives are reported. As expected, MCC and F score are comparable to each other and increase as effect size increases (**Supplementary Figure S2**). This consistent pattern is observed across four scenarios for ARZIMM but not for Poisson nor MAR models. Similar to TPR and FPR estimates, the MCC and F score values of Poisson and ARZIMM models are coincident under Scenario 1. However, in other situations, both Poisson and MAR perform poorly with low MCC and F score values.

As for the computational cost, ARIZMM took about 2.4 h to complete the estimation and bootstrap inference for a simulated dataset with 50 subjects, 20 timepoints, and 20 taxa.

Real Data Application

We applied ARZIMM methods to the MIME study. The MIME study is an ongoing randomized trial on 80 healthy volunteers with one control group (ctrl) and two antibiotic groups (amoxicillin, amx, and azithromycin, azm); antibiotics are provided for a 1-week period at the start of the trial. The main microbiome research goal of the MIME study is to evaluate the effects of antibiotics on microbial profiles at both the community and taxonomical levels. With ARIZMM, we propose a different perspective to evaluate the effect of antibiotics through the investigation of microbial interaction and community stability across groups. Because the clinical trial is still ongoing and only partial data are available, the following data analysis is done on a subset of MIME data including only 11 subjects who were randomized to two groups: 4 ctrls and 7 azms. The main purpose of this analysis is to illustrate how to use ARIZMM, not for the scientific conclusion. For each subject, we collected two baseline microbiome samples, three samples during the course of antibiotics, and five post-antibiotic samples. The gut microbiota of these individuals were profiled using 16S rRNA gene targeted sequencing on the Illumina MiSeq platform. To obtain the microbial absolute abundances, we multiplied the relative abundances of OTUs by the sample density 1.1 g/cm^3 and the number of universal 16S rRNA per gram measured using qPCR (Stein et al., 2013a). In our analysis, samples that collected before treatment in both antibiotic groups were excluded. The abundances of taxa were agglomerated at the genus level and taxa were further filtered if 1) the average relative abundances over all samples are less than 0.1%, and 2) the taxa are presented in less than 5 samples within each group.

First, **Figure 4A** shows a comparison of the relative abundance (top panel) and the absolute abundances determined by quantitative sequencing (bottom panel) of the dominant bacterial genera in 99 fecal samples from 11 subjects (blocks) across seven to nine time points (shown from left to right within each block) of this preliminary dataset. It is evident that the relative abundance and absolute abundance data present different information about the microbial profiles, and that the total bacterial load changes over time for each subject (i.e., within each block). Thus it is essential to study the microbial interactions using the absolute abundance data.

Then, we evaluate the model fitting of the log-normal distribution [used in MAR(1)] and zero-inflated over-dispersed Poisson distribution (used in ARZIMM) on the available subset of MIME data using chi-square goodness of fit test at 5% significance level taxon by taxon. Out of 45 taxa in the control group, 1 and 44 of their absolute abundances were fitted well ($p > 0.05$) by log-normal distribution and zero-inflated over-dispersed Poisson distribution respectively. The log-normal distribution fails to fit the data well when microbial taxa's absolute abundance data are left-skewed and sparse (two examples are illustrated in **Figure 4B**).

Next we demonstrate how to conduct inference for microbial interactions and community stability with ARZIMM on MIME data. First, we fit ARZIMM to ctrl and azm groups separately, adjusting for age, gender, and BMI, to get their estimated interaction matrix $\hat{\boldsymbol{B}}$ s. **Table 2** reports the characteristics of microbial interaction matrix estimates $\hat{\boldsymbol{B}}$ s. Defining the interaction effect as informative if its $\hat{B}_{mj}$'s 95% bootstrap confidence interval (based on 100 bootstrap samples) does not contain zero, we identified 125 and 105 informative interactions, respectively, in azm and ctrl groups. Their interaction effects are illustrated using networks in **Figure 5**. With more informative interactions, the azm groups have bigger and more complex networks than the ctrl group (first row of **Figure 5**), while the control group has more large estimated interaction effects than those in azm group as showed in **Table 2** and the last three rows of **Figure 5**. This observation indicates that the antibiotic treatment reduce the strength of the interactions among the taxa and create more variations with more weak interactions among taxa, thus reduce its stability. In the last row of **Table 2**, based on our stability theory we report the stability properties of the studied microbial communities. The ctrl group has the lower estimates of maximum eigenvalue squared 0.11 comparing to the azm group's maximum eigenvalue squared 0.32, which indicates that the control microbial community is more stable than the antibiotic communities.

Figure 6 provides additional information on the network feature comparison between ctrl and azm groups. **Figure 6A** displays the distribution of the positive and negative informative interaction estimates separately. The ratios between the numbers of positive and negative interactions are both around 1:1 in two groups. **Figure 6B** presents the frequency distribution of vertex degree of all the taxa in each group and they are all skew to the right. In the figure, a vertex represents a taxon in a community and its vertex degree is the number of informative interaction effect it has with the other taxa. By defining average neighbor degree as the average number of a given taxon's neighbor vertices' degrees, **Figure 6C** shows that the average neighbor degree is negatively correlated with the vertex degree in azm antibiotic treated group, but not in the control group. This indicates that there may be a group of taxa interacting with each other actively in the antibiotic group. It would be interesting to identify such sub-community with additional effort.

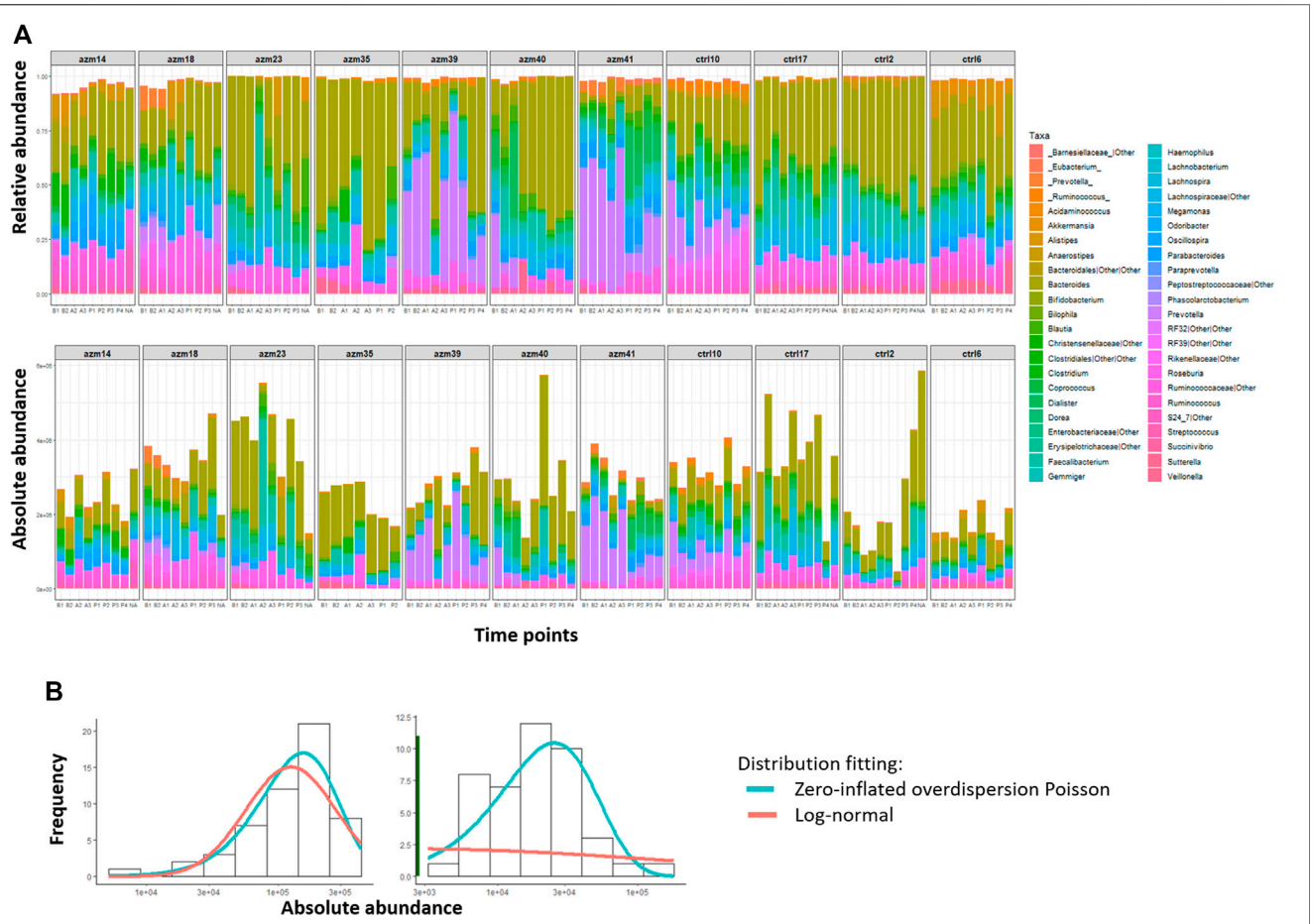


FIGURE 4 | MIME study microbiome data. **(A)** Difference between relative abundances (top panel) and absolute abundances based on qPCR (bottom panel) of dominant genera in XX fecal samples obtained from 21 subjects (block) at 7–9 time points (*x*-axis) each. **(B)** Distribution of absolute abundances of two representative genera from the MIME study, shown in the left and right panels, respectively. For each genus, the absolute abundance is fitted with a log-normal distribution (red line) or a two-part distribution: a zero part (dark green line shown in right panel) and a non-zero part fitted with an over-dispersion Poisson distribution (blue line).

TABLE 2 | The characteristics of networks.

Group description	Azithromycin	Control
Sample size	7	4
Number of time points	9	9
Number of taxa	49	45
Number of informative interactions	125	105
Number of $\lvert\hat{B}_{mj}\rvert < 0.1$	73	45
Number of $0.1 \leq \lvert\hat{B}_{mj}\rvert < 0.25$	30	29
Number of $0.25 \leq \lvert\hat{B}_{mj}\rvert < 0.5$	17	14
Number of $\lvert\hat{B}_{mj}\rvert \geq 0.5$	5	17
Informative interaction percentage (%)	5.21	5.19
Maximum eigenvalue squared	0.32	0.11

DISCUSSION

In this paper, we propose ARZIMM, an analytic platform which estimates the microbial interactions and community stability using longitudinal microbiome data. ARZIMM tackles the zero-inflated absolute abundance with a mixture distribution of zero and exponential dispersion distribution family, and enhances statistical efficiency by utilizing a random-effects term to account for the correlations among repeated measurements.

It is well-known that microbial correlations calculated from relative abundances are distorted by the compositional nature of microbiome data, and are insufficient in tracking microbial dynamics(Gloor et al., 2017). We advocate to investigate the microbial correlations using longitudinal absolute abundances which can be determined by combining gene amplicon sequencing with auxiliary total DNA quantitation data. qPCR is one of the most commonly used strategies to quantify total DNA (Dannemiller et al., 2014) and has been implemented in various statistical analyses (Stein et al., 2013b). Other alternative methods to quantify the absolute abundances include the combination of the sequencing approach (16S rRNA gene) with robust single-cell enumeration technologies (flow cytometry) (Props et al., 2017) and the usage of synthetic chimeric DNA spikes (Tkacz et al., 2018b).

Plenty of zero-inflated mixed effects models have been recently proposed to handle the excess zeros in microbiome abundance data such as zero-inflated Poisson, negative

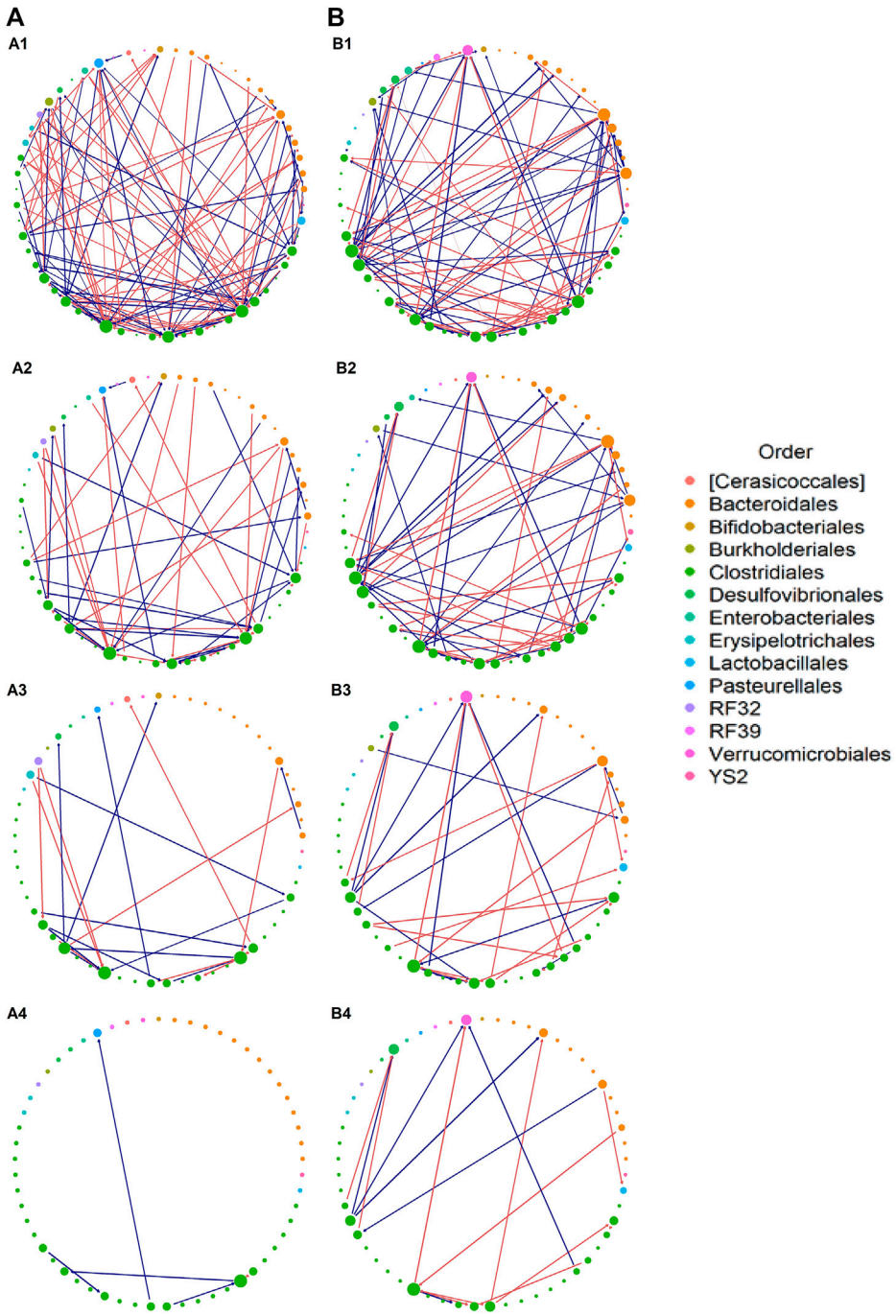


FIGURE 5 | Interaction network. Estimated interaction network for: **(A)** azithromycin (azm), and **(B)** control groups, displaying **(1)** all selected interactions, **(2)** interactions with $|\hat{B}_{mj}| \geq 0.1$, **(3)** interactions with $|\hat{B}_{mj}| \geq 0.25$, and **(4)** interactions with $|\hat{B}_{mj}| \geq 0.5$. Each node represents a taxon at the genus level, the size of which shows the degree of that taxa and the color of which shows the phylogenetic Order level for each taxon. Each edge with arrow represents an interaction effect, the width of which represents the absolute effect size on a $\log_{10}$ scale, with the color showing a positive (orange) or negative (blue) effect.

binomial and quasi-Poisson models(Xia et al., 2018; Zhang et al., 2018). However, none of the existing methods estimates the microbial interactions and community stability. To fill this gap, we extended a zero-inflated Poisson model with auto-regression and random effects modeling, which plays crucial role in efficiently handling the individual heterogeneity and enable the investigation of microbial interactions.

We investigated two community stability measurements derived from ARZIMM: the return rate and reactivity, to further understand ecological dynamics. The estimated

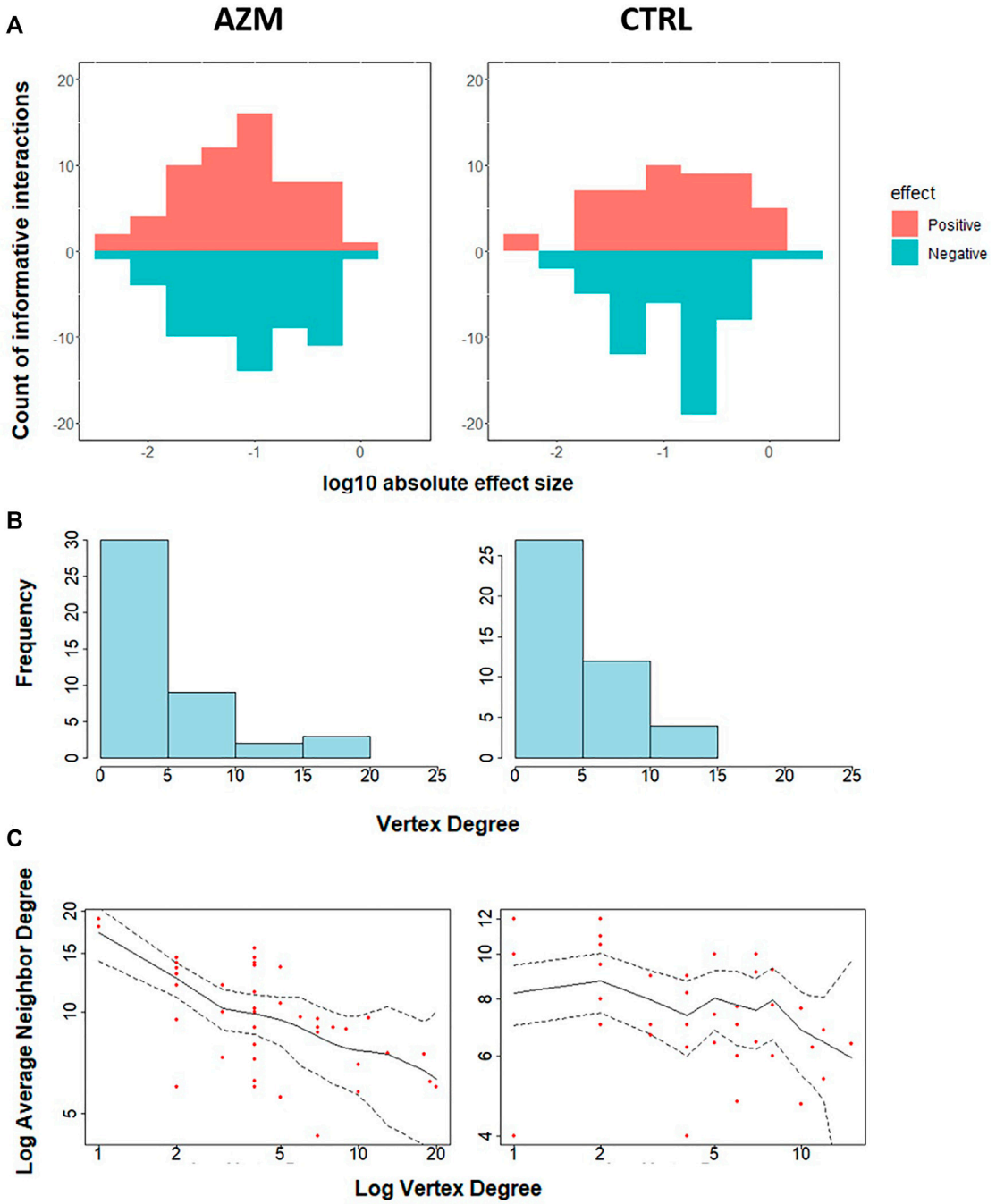


FIGURE 6 | Characteristics of estimated interactions. **(A)** The effect size of estimated informative interactions, wherein the *x*-axis represents the $\log_{10}$ scaled absolute effect size, the *y*-axis represents the count of informative interactions, and the colors represent the positive or negative effects. **(B)** Histogram of vertex degree, wherein given a vertex, vertex degree is defined as the counts of edges upon the vertex. **(C)** The average neighbor degree (*y*-axis) versus vertex degree on a log-log scale (*x*-axis). The average neighbor degree is the average number of a given taxon's neighbor vertices' degrees. Dotted lines represent 95% confidence limits.

interaction matrix $\boldsymbol{B}$ from the ARZIMM model serves the basis to calculate the largest eigenvalue of $\boldsymbol{B}$: $\max(\lambda_{\boldsymbol{B}})$, which determines the return rate of the mean of the transition distribution from the departure to the mean of the stationary distribution. We proposed to measure the reactivity of a microbial community by the expected change of the stationary distribution's mean in distance from one time point to the next time point. In ARZIMM, higher reactivity coincides with larger eigenvalues of $\boldsymbol{B}$, thus governed again by $\max(\lambda_{\boldsymbol{B}})$. Other measures of community stability, such as variance of the stationary distribution (Ives et al., 2003), warrant further investigations.

It is worth noting that by utilizing the ARZIMM model framework, the time-dependent perturbation (for instance, diet) can also be assessed flexibly in both the autoregressive part and the logistic part in the model. However, the stability based on the microbial interactions has to be interpreted with caution, since the mean of stationary distribution changes along with the time-dependent covariates.

We have demonstrated that ARZIMM outperforms the competing methods and exhibits its feasibility for examining microbial interactions and stability based on longitudinal microbial data. We applied our method to a real human microbiome study of antibiotic treatment and elucidated the microbial interaction network of bacteria from antibiotic and non-antibiotic groups separately. The application of ARZIMM to temporal microbiome data shows great promise. Still, the development of accurate predictive models will require further developments. For example, the method used here to infer microbial interactions may be expanded by adding functional information as well as phylogenetic information. Although this method is primarily developed for the gut microbiota, it may be potentially applied to longitudinal data from any ecological systems. Since interactions between members of microbial communities are primary driving forces for the long-term stability(Ratzke et al., 2020), the corresponding stability properties will provide useful principles for community dynamics.

Note that the proposed ARIZMM assumes the probability of observing a zero count for a taxon is constant over time. The reason is two-fold. 1) One major goal of ARIZMM is to derive the inference on the stability of the microbial community over a certain period. With the constant probability of observing a zero count assumption, the stability inference will solely depend on the estimation of the taxon-by-taxon interaction matrix $\boldsymbol{B}$. Otherwise, a stationary distribution will not exit. 2) Using the MIME data, we estimated the proportions of zeros (denoted as q_{mt}) for all taxa by group at all time points, then calculated the mean($\overline{\hat{q}}_m$) and standard deviation ($SD_{\hat{q}_m}$) over all the time points and the coefficient of variation ($CV_m = SD_{\hat{q}_m}/\overline{\hat{q}}_m$, $m = 1, \ldots, M$) to evaluate their temporal variations. The median of CV_m over all taxa in the control, Amoxicillin and Azithromycin groups are 0.16, 0.12, and 0.34 respectively. This results reveal two observations: 1) the temporal variations of q_{mt} in most taxa are relative weak; and 2) the temporal variation of the proportions of zero is heterogeneous and there may be no one perfect model fitting all the taxa well. Thus, we believe our assumption that p_m is constant over time is valid and pragmatic. To further check the robustness of our proposed model, we conducted additional simulation by introducing extra randomness when we generate the probability of observing a zero count across the time points, while analyze the data using our proposed model. Our results show that the moderate temporal variation in probability of zero count does not affect ARIZMM's performance much in capturing the informative interactions by estimating $\boldsymbol{B}$ when the absolute effect strengths of interaction matrices is high or medium. The detailed simulation design and results are reported in the **Supplementary Material Section S4.2** and **Supplementary Figure S3**.

The proposed method, ARZIMM has a few limitations and future works are needed to improve it. ARZIMM adopts a simple correlation structure that the random effects in the multivariate logistic component and the multivariate autoregressive component $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are assumed independent. We took this parsimonious model based on our experience(Hu, 2021; Wang, 2021) in modeling the longitudinal microbiome data to ease the computational burden. The more general random effects structure with cross-part correlations can provide more robust modeling, however, can suffer from model convergence as well. Further investigation is warranted.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The motivating study is still ongoing. After it is closed, it will be uploaded to the public repository. Requests to access these datasets should be directed to martin.blaser@cabm.rutgers.edu.

AUTHOR CONTRIBUTIONS

LH and HL developed the methodological ideas. LH implemented the methods, performed the simulations and real data analysis, and developed the software package. CW and JH contributed to the simulation design and real data analysis. ZG, EF, SH, and MB contributed to the acquisition of utilized real microbiome data. MB provided biological insights and interpretation of the real data analysis. LH and HL wrote the manuscript. All authors read, edited, and approved the final manuscript.

FUNDING

This work was supported in part by National Institutes of Health grants R01DK110014, P20CA252728, and U01AI22285.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.777877/full#supplementary-material

REFERENCES

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., et al. (2016). MDSINE: Microbial Dynamical Systems INference Engine for Microbiome Time-Series Analyses. *Genome Biol.* 17 (1), 121–217. doi:10.1186/s13059-016-0980-6

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., et al. (2016). MDSINE: Microbial Dynamical Systems INference Engine for Microbiome Time-Series Analyses. *Genome Biol.* 17 (1), 121. doi:10.1186/s13059-016-0980-6

Carpenter, S. R., Cole, J. J., Pace, M. L., Batt, R., Brock, W. A., Cline, T., et al. (2011). Early Warnings of Regime Shifts: a Whole-Ecosystem experiment. *Science* 332 (6033), 1079–1082. doi:10.1126/science.1203672

Carroll, I. M., Ringel-Kulka, T., Siddle, J. P., and Ringel, Y. (2012). Alterations in Composition and Diversity of the Intestinal Microbiota in Patients with Diarrhea-Predominant Irritable Bowel Syndrome. *Neurogastroenterology Motil.* 24 (6), 521–e248. doi:10.1111/j.1365-2982.2012.01891.x

Carroll, S. S., and Cressie, N. (1997). Spatial Modeling of Snow Water Equivalent Using Covariances Estimated from Spatial and Geomorphic Attributes. *J. Hydrol.* 190 (1-2), 42–59. doi:10.1016/s0022-1694(96)03062-4

Chamberlain, G. (1982). Multivariate Regression Models for Panel Data. *J. Econom.* 18 (1), 5–46. doi:10.1016/0304-4076(82)90094-x

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., et al. (2012). Gut Microbiota Composition Correlates with Diet and Health in the Elderly. *Nature* 488 (7410), 178–184. doi:10.1038/nature11319

Czado, C., Gneiting, T., and Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics* 65 (4), 1254–1261. doi:10.1111/j.1541-0420.2009.01191.x

Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic Models of the Complex Microbial Metapopulation of lake mendota. *NPJ Syst. Biol. Appl.* 2 (1), 16007–7. doi:10.1038/npjsba.2016.7

Dannemiller, K. C., Lang-Yona, N., Yamamoto, N., Rudich, Y., and Peccia, J. (2014). Combining Real-Time PCR and Next-Generation DNA Sequencing to Provide Quantitative Comparisons of Fungal Aerosol Populations. *Atmos. Environ.* 84, 113–121. doi:10.1016/j.atmosenv.2013.11.036

de Vos, M. G. J., Zagorski, M., McNally, A., and Bollenbach, T. (2017). Interaction Networks, Ecological Stability, and Collective Antibiotic Tolerance in Polymicrobial Infections. *Proc. Natl. Acad. Sci. USA* 114 (40), 10666–10671. doi:10.1073/pnas.1713372114

Douc, R., Doukhan, P., and Moulines, E. (2013). Ergodicity of Observation-Driven Time Series Models and Consistency of the Maximum Likelihood Estimator. *Stochastic Process. their Appl.* 123 (7), 2620–2647. doi:10.1016/j.spa.2013.04.010

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The Long-Term Stability of the Human Gut Microbiota. *Science* 341 (6141), 1237439. doi:10.1126/science.1237439

Faust, K., and Raes, J. (2012). Microbial Interactions: from Networks to Models. *Nat. Rev. Microbiol.* 10 (8), 538–550. doi:10.1038/nrmicro2832

Fisher, C. K., and Mehta, P. (2014). Identifying keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PloS one* 9 (7), e102451. doi:10.1371/journal.pone.0102451

Fokianos, K., and Tjøstheim, D. (2011). Log-linear Poisson Autoregression. *J. Multivariate Anal.* 102 (3), 563–578. doi:10.1016/j.jmva.2010.11.002

Gerber, G. K. (2014). The Dynamic Microbiome. *FEBS Lett.* 588 (22), 4131–4139. doi:10.1016/j.febslet.2014.02.037

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current Understanding of the Human Microbiome. *Nat. Med.* 24 (4), 392–400. doi:10.1038/nm.4517

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8, 2224. doi:10.3389/fmicb.2017.02224

Hu, J. (2021). Joint Modeling of Zero-Inflated Longitudinal Proportions and Time-To-Event Data with Application to a Gut Microbiome Study. *Biometrics* 2, 2020. doi:10.1111/biom.13515

Ives, A. R., Dennis, B., Cottingham, K. L., and Carpenter, S. R. (2003). Estimating Community Stability and Ecological Interactions from Time-Series Data. *Ecol. Monogr.* 73 (2), 301–330. doi:10.1890/0012-9615(2003)073[0301:ecsaei]2.0.co;2

Ives, A. R., Klug, J. L., and Gross, K. (2000). Stability and Species Richness in Complex Communities. *Ecol. Lett.* 3 (5), 399–411. doi:10.1046/j.1461-0248.2000.00144.x

Jackson, M. A., Jeffery, I. B., Beaumont, M., Bell, J. T., Clark, A. G., Ley, R. E., et al. (2016). Signatures of Early Frailty in the Gut Microbiota. *Genome Med.* 8 (1), 8. doi:10.1186/s13073-016-0262-7

Kim, J., Lim, J., and Lee, C. (2013). Quantitative Real-Time PCR Approaches for Microbial Community Studies in Wastewater Treatment Systems: Applications and Considerations. *Biotechnol. Adv.* 31 (8), 1358–1373. doi:10.1016/j.biotechadv.2013.05.010

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best Practices for Analysing Microbiomes. *Nat. Rev. Microbiol.* 16 (7), 410–422. doi:10.1038/s41579-018-0029-9

Liesenfeld, R., Nolte, I., and Pohlmeier, W. (2006). Modelling Financial Transaction price Movements: a Dynamic Integer Count Data Model. *Empirical Econ.* 30 (4), 795–825. doi:10.1007/s00181-005-0001-1

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., and Bar-Joseph, Z. (2019). Dynamic Interaction Network Inference from Longitudinal Microbiome Data. *Microbiome* 7 (1), 54–14. doi:10.1186/s40168-019-0660-3

Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. (2014). Mathematical Modeling of Primary Succession of Murine Intestinal Microbiota. *Proc. Natl. Acad. Sci. USA* 111 (1), 439–444. doi:10.1073/pnas.1311322111

Martinez, C., Antolin, M., Santos, J., Torrejon, A., Casellas, F., Borruel, N., et al. (2008). Unstable Composition of the Fecal Microbiota in Ulcerative Colitis during Clinical Remission. *Am. J. Gastroenterol.* 103 (3), 643–648. doi:10.1111/j.1572-0241.2007.01592.x

Maukonen, J., Satokari, R., Mättö, J., Söderlund, H., Mattila-Sandholm, T., and Saarela, M. (2006). Prevalence and Temporal Stability of Selected Clostridial Groups in Irritable Bowel Syndrome in Relation to Predominant Faecal Bacteria. *J. Med. Microbiol.* 55 (5), 625–633. doi:10.1099/jmm.0.46134-0

McGeachie, M. J. (2016). Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Scientific Rep.* 6 (1), 1–11. doi:10.1038/srep20359

Mounier, J., Monnet, C., Jacques, N., Antoinette, A., and Irlinger, F. (2009). Assessment of the Microbial Diversity at the Surface of Livarot Cheese Using Culture-dependent and Independent Approaches. *Int. J. Food Microbiol.* 133 (1-2), 31–37. doi:10.1016/j.ijfoodmicro.2009.04.020

Nadkarni, M. A., Martin, F. E., Jacques, N. A., and Hunter, N. (2002). Determination of Bacterial Load by Real-Time PCR Using a Broad-Range (Universal) Probe and Primers Set. *Microbiology (Reading)* 148 (Pt 1), 257–266. doi:10.1099/00221287-148-1-257

Ott, S. J., Musfeldt, M., Ullmann, U., Hampe, J., and Schreiber, S. (2004). Quantification of Intestinal Bacterial Populations by Real-Time PCR with a Universal Primer Set and Minor Groove Binder Probes: a Global Approach to the Enteric flora. *J. Clin. Microbiol.* 42 (6), 2566–2572. doi:10.1128/jcm.42.6.2566-2572.2004

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., et al. (2017). Absolute Quantification of Microbial Taxon Abundances. *Isme J.* 11 (2), 584–587. doi:10.1038/ismej.2016.117

Ratzke, C., Barrere, J., and Gore, J. (2020) Strength of Species Interactions Determines Biodiversity and Stability in Microbial Communities. *Nat. Ecol. Evol.* 4, 376–383. doi:10.1038/s41559-020-1099-4

Russell, S., and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: University of California at Berkeley.

Scanlan, P. D., Shanahan, F., Clune, Y., Collins, J. K., O'Sullivan, G. C., O'Riordan, M., et al. (2008). Culture-independent Analysis of the Gut Microbiota in Colorectal Cancer and Polyposis. *Environ. Microbiol.* 10 (3), 789–798. doi:10.1111/j.1462-2920.2007.01503.x

Shade, A., Gregory Caporaso, J., Handelsman, J., Knight, R., and Fierer, N. (2013). A Meta-Analysis of Changes in Bacterial and Archaeal Communities with Time. *Isme J.* 7 (8), 1493–1506. doi:10.1038/ismej.2013.54

Shankar, J. (2017). Insights into Study Design and Statistical Analyses in Translational Microbiome Studies. *Ann. Transl. Med.* 5 (12), 249. doi:10.21037/atm.2017.01.13

Shaw, G. T., Pao, Y. Y., and Wang, D. (2016). MetaMIS: a Metagenomic Microbial Interaction Simulator Based on Microbial Community Profiles. *BMC bioinformatics* 17 (1), 488–512. doi:10.1186/s12859-016-1359-0

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *Plos Comput. Biol.* 9 (12), e1003388. doi:10.1371/journal.pcbi.1003388

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *Plos Comput. Biol.* 9 (12), e1003388. doi:10.1371/journal.pcbi.1003388

Tkacz, A., Hortala, M., and Poole, P. S. (2018). Absolute Quantitation of Microbiota Abundance in Environmental Samples. *Microbiome* 6 (1), 110–113. doi:10.1186/s40168-018-0491-7

Tkacz, A., Hortala, M., and Poole, P. S. (2018). Absolute Quantitation of Microbiota Abundance in Environmental Samples. *Microbiome* 6 (1), 110. doi:10.1186/s40168-018-0491-7

Uronis, J. M., Mühlbauer, M., Herfarth, H. H., Rubinas, T. C., Jones, G. S., and Jobin, C. (2009). Modulation of the Intestinal Microbiota Alters Colitis-Associated Colorectal Cancer Susceptibility. *PloS one* 4 (6), e6026. doi:10.1371/journal.pone.0006026

Venturelli, O. S., Carr, A. C., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., et al. (2018). Deciphering Microbial Interactions in Synthetic Human Gut Microbiome Communities. *Mol. Syst. Biol.* 14 (6), e8157. doi:10.15252/msb.20178157

Ver Hoef, J. M., and Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data. *Ecology* 88 (11), 2766–2772. doi:10.1890/07-0043.1

Wang, C. (2021). Microbial Trend Analysis for Common Dynamic Trend, Group Comparison and Classification in Longitudinal Microbiome Study. *BMC Genomics* 15, 667. doi:10.1186/s12864-021-07948-w

Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., and Yuen, K.-Y. (2008). Then and Now: Use of 16S rDNA Gene Sequencing for Bacterial Identification and Discovery of Novel Bacteria in Clinical Microbiology Laboratories. *Clin. Microbiol. Infect.* 14 (10), 908–934. doi:10.1111/j.1469-0691.2008.02070.x

Xia, Y., Sun, J., and Chen, D.-G. (2018). *Statistical Analysis of Microbiome Data with R*, 847. Springer.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., et al. (2018). Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. *Front. Microbiol.* 9, 1683. doi:10.3389/fmicb.2018.01683

Zuo, T., and Ng, S. C. (2018). The Gut Microbiota in the Pathogenesis and Therapeutics of Inflammatory Bowel Disease. *Front. Microbiol.* 9, 2247. doi:10.3389/fmicb.2018.02247

Conflict of Interest: LH was employed by the company Novartis Pharmaceuticals Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2022
doi: 10.3389/fgene.2022.799615

Incorporation of Data From Multiple Hypervariable Regions when Analyzing Bacterial 16S rRNA Gene Sequencing Data

Carli B. Jones[1†], James R. White[2], Sarah E. Ernst[1], Karen S. Sfanos[1,3,4] and Lauren B. Peiffer[1,5*†]*

[1]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Resphera Biosciences, Baltimore, MD, United States, [3]Deparment of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [4]Department of Urology, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [5]Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

OPEN ACCESS

Edited by:
Himel Mallick,
Merck, United States

Reviewed by:
Jonathan Badger,
National Cancer Institute,
United States
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

***Correspondence:**
Lauren B. Peiffer
lpeiffe1@jhmi.edu
Karen S. Sfanos
ksfanos@jhmi.edu

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 21 October 2021
Accepted: 08 March 2022
Published: 31 March 2022

Citation:
Jones CB, White JR, Ernst SE, Sfanos KS and Peiffer LB (2022) Incorporation of Data From Multiple Hypervariable Regions when Analyzing Bacterial 16S rRNA Gene Sequencing Data. Front. Genet. 13:799615. doi: 10.3389/fgene.2022.799615

Short read 16 S rRNA amplicon sequencing is a common technique used in microbiome research. However, inaccuracies in estimated bacterial community composition can occur due to amplification bias of the targeted hypervariable region. A potential solution is to sequence and assess multiple hypervariable regions in tandem, yet there is currently no consensus as to the appropriate method for analyzing this data. Additionally, there are many sequence analysis resources for data produced from the Illumina platform, but fewer open-source options available for data from the Ion Torrent platform. Herein, we present an analysis pipeline using open-source analysis platforms that integrates data from multiple hypervariable regions and is compatible with data produced from the Ion Torrent platform. We used the ThermoFisher Ion 16 S Metagenomics Kit and a mock community of twenty bacterial strains to assess taxonomic classification of six amplicons from separate hypervariable regions (V2, V3, V4, V6-7, V8, V9) using our analysis pipeline. We report that different amplicons have different specificities for taxonomic classification, which also has implications for global level analyses such as alpha and beta diversity. Finally, we utilize a generalized linear modeling approach to statistically integrate the results from multiple hypervariable regions and apply this methodology to data from a representative clinical cohort. We conclude that examining sequencing results across multiple hypervariable regions provides more taxonomic information than sequencing across a single region. The data across multiple hypervariable regions can be combined using generalized linear models to enhance the statistical evaluation of overall differences in community structure and relatedness among sample groups.

Keywords: 16S rRNA, microbiome, hypervariable regions, sequencing, ion torrent

INTRODUCTION

Next generation sequencing of microbial DNA has become an important tool used for determining relationships between human-associated microbial populations and various diseases. Most studies in this realm rely on either shotgun metagenomic sequencing or 16 S ribosomal RNA (rRNA) amplicon sequencing. Shotgun metagenomic sequencing involves sequencing random fragments of sample DNA which contains a mixture of bacterial DNA, as well as host and other microbial and environmental DNA (Quince et al., 2017). This method allows for taxonomic profiling, metabolic function profiling, and antibiotic resistance gene profiling; however, it is generally more expensive than amplicon sequencing, and requires a larger amount of input DNA and the availability of reference genome sequences. Bacterial 16 S rRNA amplicon sequencing employs PCR amplification of specific hypervariable regions within the gene, followed by deep sequencing (Sanschagrin and Yergeau, 2014). This method is generally a quicker, cheaper alternative to shotgun metagenomics; however, it only identifies bacteria and the typical strategy only sequences a specific fragment of the bacterial 16 S rRNA gene (Ranjan et al., 2016). While functional information can be inferred from taxonomic classification using tools such as UniRef and KEGG Orthology, the genetic elements contributing to these functions themselves are not sequenced. The 16 S rRNA gene is comprised of 9 hypervariable regions (V1-V9), and most primers used for next generation sequencing only target one to two hypervariable regions at a time. Multiple studies have shown that different regions vary in their taxonomic utility due to a combination of primer bias, differential hypervariable region sequence length, and hypervariable region sequence uniqueness across bacterial taxa (Claesson et al., 2010; Pinto and Raskin, 2012; Cai et al., 2013; Tremblay et al., 2015; Barb et al., 2016). An ideal solution would be to sequence the entire 16 S rRNA gene, however this technique is more costly and access to this technology is limited compared to traditional 16 S rRNA sequencing. Therefore, a potential alternative would be to perform 16 S rRNA amplicon sequencing on multiple regions and incorporate information from as many hypervariable regions as possible into downstream data analysis.

The Ion 16 S™ Metagenomics Kit (Life Technologies) utilizes six sets of primers spanning seven different hypervariable regions: V2, V3, V4, V6-7, V8, and V9. This is an attractive approach because it yields more sequence information across the 16 S rRNA gene overall. However, there is currently little consensus as to how to properly analyze information from multiple hypervariable regions and obtain overall results. Current analysis pipelines for Ion Torrent data include the Ion Reporter Software offered by ThermoFisher, and an alternative method using open access tools developed by Barb *et al.* (Barb et al., 2016). The utility of Ion Reporter Software is limited; for example, users are unable to incorporate study-specific metadata into analyses, and exported processed data is devoid of previous analysis information, preventing downstream analysis with open-source tools. Barb *et al.* offer methods for taxonomic identification; however, they do not address the question of how to appropriately integrate data from multiple hypervariable regions in downstream analyses. Recently, Fuks *et al.* (Fuks et al., 2018) and Debelius *et al.* (Debelius et al., 2021) developed methods to computationally combine data from multiple hypervariable regions to provide a joint estimate of the microbial community composition. To date, however, there is no generally agreed upon approach for combining sequences from multiple hypervariable regions for downstream analyses, especially for less commonly used 16 S rRNA gene sequencing platforms such as Ion Torrent.

Herein, we developed an analysis pipeline that analyzes data from each hypervariable region separately, allowing for systematic comparison of taxonomic classification by hypervariable region. We demonstrate our results from analyzing a mock community of bacterial DNA where we determine how each hypervariable region differs in its utility to provide information on taxonomic classifications, alpha diversity, and beta diversity. We report that certain taxa are only identified by particular hypervariable regions, corroborating prior studies (Claesson et al., 2010; Pinto and Raskin, 2012; Cai et al., 2013; Tremblay et al., 2015; Barb et al., 2016) and supporting our hypothesis that there is a benefit to incorporating multiple primer sets into sequencing strategies. Furthermore, we discuss different options for downstream analysis and statistics, and demonstrate that using a generalized linear model (GLM) to statistically combine results from multiple hypervariable regions increases sensitivity of taxonomic classification. Finally, we demonstrate the utility of our approach in the analysis of clinical samples in an illustrative clinical cohort.

MATERIALS AND METHODS

Mock Community

The 20 Strain Even Mix Genomic Material was obtained from American Type Culture Collection (ATCC, Cat. No. MSA-1002, Manassas, VA). The strain composition of the mock community is given in **Table 1**. The mock community was sequenced a total of five times from four library preparations and over three sequencing runs.

Clinical Sample Collection

All specimens were studied under an Institutional Review Board (IRB) approved protocol with written informed consent. A total of three (3) adult males self-collected two (2) rectal swab samples each with sterile flocked swabs (Cat. No. 552C, Copan Diagnostics, Murrieta, CA). One rectal swab from each individual was randomly selected for DNA extraction immediately after sample collection (RS1). The other swab (RS2) was frozen at −80°C for 6 days before DNA extraction.

DNA Extraction

The DNA extraction protocol was adapted from our previously published protocol (Shrestha et al., 2018). Briefly, rectal swab fecal material was resuspended in 500 µl of 1X phosphate buffered saline (PBS) (Cat. No. 21-031-CV, Corning, Manassas, VA). Samples were then digested in a cocktail of lysozyme (10 mg/

TABLE 1 | Contents of mock community.

Species	16S copies[a]	Genus	Family
Acinetobacter baumannii	6	*Acinetobacter*	*Moraxellaceae*
Actinomyces odontolyticus	2	*Actinomyces*	*Actinomycetaceae*
Bacillus cereus	12	*Bacillus*	*Bacillaceae*
Bacteroides vulgatus	7	*Bacteroides*	*Bacteroidaceae*
Bifidobacterium adolescentis	5	*Bifidobacterium*	*Bifidobacteriaceae*
Clostridium beijerinckii	14	*Clostridium*	*Clostridiaceae*
Cutibacterium acnes	4	*Cutibacterium*	*Propionibacteriaceae*
Deinococcus radiodurans	7	*Deinococcus*	*Deinococcaceae*
Enterococcus faecalis	4	*Enterococcus*	*Enterococcaceae*
Escherichia coli	7	*Escherichia*	*Enterobacteriaceae*
Helicobacter pylori	2	*Helicobacter*	*Helicobacteraceae*
Lactobacillus gasseri	6	*Lactobacillus*	*Lactobacillaceae*
Neisseria meningitidis	4	*Neisseria*	*Neisseriaceae*
Porphyromonas gingivalis	4	*Porphorymonas*	*Porphyromonadaceae*
Pseudomonas aeruginosa	4	*Pseudomonas*	*Pseudomonadaceae*
Rhodobacter sphaeroides	3	*Rhodobacter*	*Rhodobacteraceae*
Staphylococcus aureus	6	*Staphylococcus*	*Staphylococcaceae*
Staphylococcus epidermidis	5	*Staphylococcus*	*Staphylococcaceae*
Streptococcus agalactiae	7	*Streptococcus*	*Streptococcaceae*
Streptococcus mutans	5	*Streptococcus*	*Streptococcaceae*

[a]*Number of copies of 16S rRNA genes contained in the bacterial genome of the indicated species.*

ml, Cat. No. L7773, Sigma-Aldrich, St. Louis, MO) and mutanolysin (25 KU/ml, Cat. No. M4782, Sigma-Aldrich, St. Louis, MO) for 1 h at 37°C. The contents of the tubes were then transferred into FastPrep Lysing Matrix B tubes (Cat. No. 6911050, MP Biomedicals, Santa Ana, CA). Next, 20% SDS (Cat. No. 05030, Sigma-Aldrich, St. Louis, MO) and phenol: chloroform:isoamyl alcohol (25:24:1, Cat. No. 108-95-2, ThermoFisher Scientific, Waltham, MA) were added and samples were homogenized by bead beating in an MP FastPrep-24 at 6 m/s for a total of 60 s. DNA was precipitated and resuspended in a final volume of 50 µl of DNA-free water (Cat. No. P-020-0003, Molzym, Bremen, Germany).

Library Preparation

Concentration of DNA from the mock microbial community (**Table 1**) and rectal swabs was measured using a Qubit dsDNA HS (high sensitivity) kit (Cat. No. Q32851, Life Technologies, Carlsbad, CA). Libraries were prepared using the Ion 16 S™ Metagenomics Kit (Cat. No. A26216, ThermoFisher Scientific, Waltham, MA). Briefly, 10 ng of DNA was mixed with 15 µl of Environmental Master Mix. 3 µl of each 16 S Primer Set (10X) was added to each tube, one sample set with primers for V2-4-8 (Pool 1) and the other with primers for V3-6,7-9 (Pool 2). Samples were placed in a thermocycler with the following thermal conditions: 95°C for 10 min; then 25 cycles of 95°C for 30 s, 58°C for 30s, 72°C for 30 s; and finally 72°C for 7 min. Amplification products were purified using AMPure XP beads (Cat. No. A63881, Beckman Coulter, Pasadena, CA) and eluted in nuclease free water. Concentrations of amplification products from Pool 1 and Pool 2 were measured using a Bioanalyzer High Sensitivity DNA Kit (Cat. No. 5067-4626, Agilent Technologies, Santa Clara, CA), and the two pools were combined for a total of 100 ng of DNA (50 ng from each pool).

Next, 20 µl of 5X End Repair Buffer and 1 µl of End Repair Enzyme were added to each sample, and then incubated for 20 min at room temperature. Pooled amplicons were then purified again using AMPure XP beads and eluted in Low TE buffer. Ligation and nick repair were performed using ×10 Ligase Buffer, Ion P1 Adaptor, Ion Xpress Barcodes, dNTP Mix, DNA Ligase, Nick Repair Polymerase, nuclease-free water, and sample DNA with the following thermal conditions: 25°C for 15 min, 72°C for 5 min. Adapter-ligated and nick-repaired DNA was then purified using AMPure XP beads and eluted in Low TE buffer.

The library was then amplified using the Ion Plus Fragment Library Kit (Cat. No. 4471252, ThermoFisher Scientific) with the following thermal conditions: 95°C for 5 min; then 7 cycles of 95°C for 15 s, 58°C for 15 s, 70°C for 1 min; and then finally 70°C for 1 min. The amplified library was then purified using AMPure XP beads and eluted in Low TE buffer. Library concentrations were measured using a Bioanalyzer and the High Sensitivity DNA Kit. Libraries were then diluted down to 26 pM and pooled, yielding a 26 pM solution.

Sequencing

Libraries were prepared for sequencing using oil amplification to template the libraries onto beads and loaded onto chips using the Ion Chef Instrument and the Ion 520™ & Ion 530™ Kit–Chef (ThermoFisher Scientific). Chips were then loaded onto the Ion GeneStudio S5 System along with Ion S5 Sequencing Kit reagents (Cat. No. A35850, ThermoFisher Scientific, Waltham, MA) and sequenced at the Sidney Kimmel Comprehensive Cancer Center Experimental and Computational Genomics Core facility. Samples in this study were sequenced across three separate sequencing runs on Ion 520 and Ion 530 chips using 400bp sequencing kits. Sequences were demultiplexed by sample using the S5 device software, and then separated per hypervariable region by ThermoFisher prior to downstream analysis.

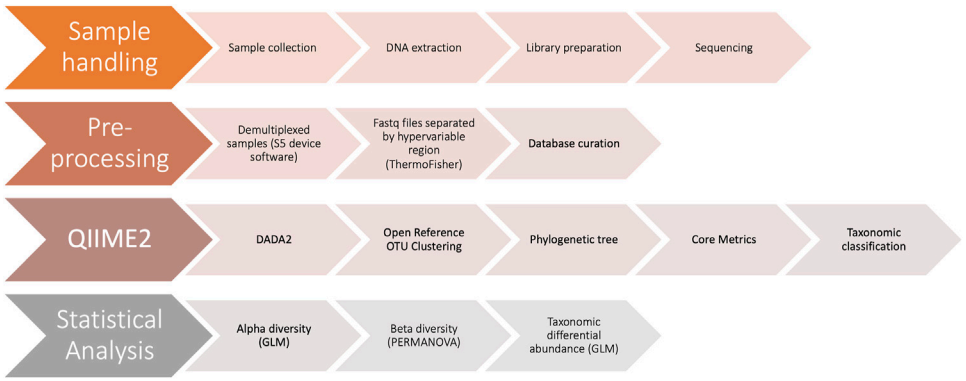


FIGURE 1 | Schematic diagram of workflow. The four major steps in our workflow include 1) sample handling, from sample collection to sequencing 2) pre-processing of sequencing data and taxonomic reference database 3) performing microbiome bioinformatics using QIIME2 and 4) statistical analysis of results using R.

Data Processing

Primer sequences are not made available to Ion 16 S™ Metagenomics Kit users. Therefore, FASTQ files had to be separated by primer set by the ThermoFisher Bioinformatics team, resulting in six separate FASTQ files per sample (V2, V3, V4, V6-7, V8, and V9), with primer sequences removed and all reads oriented in the forward direction.

Manifest files were then created for each hypervariable region and each sequencing run. FASTQ files were imported into QIIME2 format *via* qiime tools import in SingleEndFastqManifestPhred33V2 format (Bolyen et al., 2019). QIIME2 v 2020.6 was used to perform denoising, Operational Taxonomic Unit (OTU) clustering, taxonomic classification, phylogenetic tree construction, and alpha and beta diversity.

DADA2 was used to denoise data, using the denoise-pyro plugin and parameters of 0 bp for trimming and truncation (Callahan et al., 2016). A separate DADA2 run was performed for each hypervariable region and each sequencing run. Denoising statistics were then summarized and exported to P03-summarize-qc and P13-summarize-qc directories in the analysis folder of the it-workflow repository for the ATCC mock community samples and the clinical samples, respectively. From these summaries, we determined that all samples in all hypervariable regions had a minimum of 10,000 reads which passed the filter in the DADA2 step. Good's coverage was performed at a depth of 10,000 reads for each hypervariable region and at least 99% coverage was achieved for all regions (Good, 1953). Thus, we decided that 10,000 reads was an acceptable sampling depth. DADA2 feature tables and representative sequence files were then merged across sequencing runs so that there was only one feature table and representative sequence file per hypervariable region.

Open-reference OTU clustering was then performed using QIIME2 plugin vsearch cluster-features-open-reference (Bokulich et al., 2018). A threshold of 99% identity was used, and sequences were clustered against reference sequences from the curated sfanos_db_v4.0 database as described below.

Alpha and Beta Diversity Analysis

A phylogenetic tree was constructed for each hypervariable region using the "representative sequences" file generated from open-reference OTU clustering *via* the QIIME phylogeny align-to-tree-mafft-fasttree plugin (Faith et al., 1987; Price et al., 2010; Katoh and Standley, 2013). Community diversity was analyzed using the core-metrics-phylogenetic plugin. Briefly, the feature table produced by open-reference OTU clustering and the phylogenetic trees constructed in the previous step were input into the core-metrics-phylogenetic plugin, which performed alpha and beta diversity analyses at a sampling depth of 10,000 reads. Alpha diversity summaries were obtained and exported for Faith's phylogenetic diversity, Shannon diversity (Shannon, 2001), evenness, and observed OTUs. Distance matrices were exported for Jaccard (Jaccard, 1908), Bray-Curtis (Sorensen, 1948), weighted UniFrac (Lozupone et al., 2007), and unweighted UniFrac (Lozupone and Knight, 2005) distances. Data was imported into Rstudio for visualization of alpha diversity metrics and principal coordinates analysis (PCoA). Taxonomic classification results from each hypervariable region were aggregated into summary tables at higher taxonomic levels (phylum through species) for downstream comparative analysis. Beta-diversity distance matrices (using the measures bray-curtis, jaccard, unweighted-unifrac, and weighted-unifrac) were based on OTU profiles and were generated for each hypervariable region separately to account for region-specific OTUs. Additionally, a multi-region beta-diversity analysis incorporated species level assignments across all hypervariable regions, followed by distance matrix calculation (Canberra, Bray-Curtis, Jaccard, Euclidean, Gower, and Kulczynski) using the vegdist command in the vegan R package.

Database Curation

It is well known that curating existing taxonomic databases can lead to improved performance (Ritari et al., 2015; Clemmons et al., 2019; Myer et al., 2020). Therefore, uncultured and unclassified sequences were removed from the SILVA (v.123) database to eliminate sequences that have no practical value in taxonomic assignment. This refined database (*sfanos-db-4.0*) contains approximately 15,000 named species.

In Silico Taxonomic Validation of Curated Database

Prior to using sfanos-db-4.0 for taxonomic classification, we verified its utility by performing *in silico* taxonomic

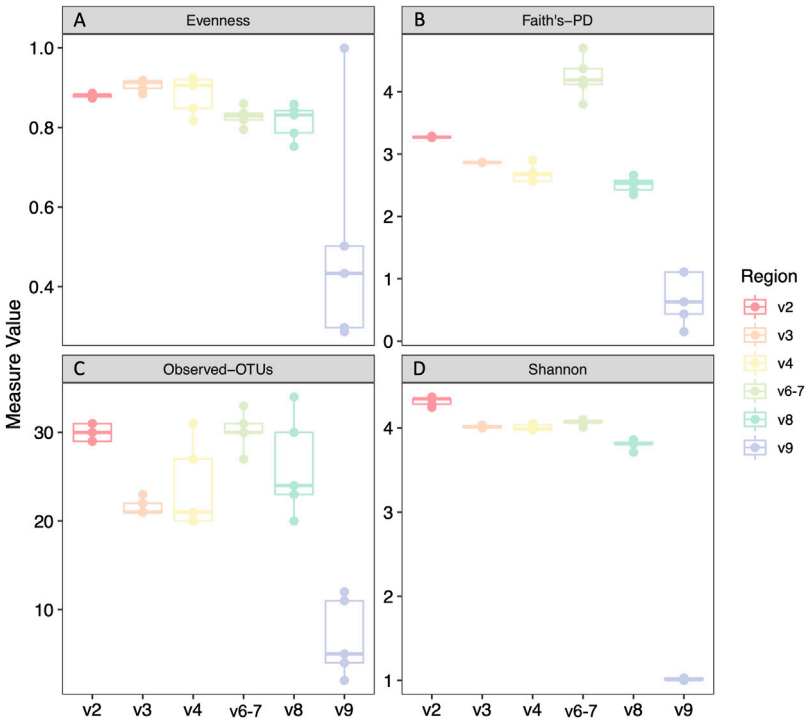


FIGURE 2 | Alpha diversity analyses of mock community technical replicates by hypervariable region. Evenness **(A)**, Faith's phylogenetic diversity **(B)**, Observed Operational Taxonomic Units (OTUs) **(C)**, Shannon diversity **(D)**. Statistical analysis and p values can be found in **Supplementary File S2**.

classification using sequences from a published human gut microbiome culture collection (Forster et al., 2019). First, we separated the sequences in the culture collection by hypervariable region to mimic our own data. To do this, we ran the sequences from the culture collection through NCBI BLAST against the ATCC mock community sequences that had already been split by hypervariable region. This method allowed us to break down the culture collection sequences into their different hypervariable regions and simulate more complex clinical data. A 1% noise rate was included in the simulated sequences to mimic typical evolutionary variation in species as well as sequencing error. We then ran taxonomic classification of the sequences from the culture collection using our curated database, with a threshold of 97% sequence identity. A confidence score was assigned to each classification by VSEARCH. Results were categorized into true positives (TP), false positives (FP), and false negatives (FN) based on whether they were found in the culture collection or not (**Supplementary File S1**). Sequence assignment counts were converted to percent by adding up the total number of sequences that were assigned as TP, FP, or FN for each V region, dividing by the total number of sequences for that region, and multiplying by 100.

Taxonomic Classification

Taxonomic classification was performed using classify-consensus-vsearch using the curated sfanos_db_v4.0 reference reads and reference taxonomy with 99% identity. The output. qza file was then exported in order to obtain the taxonomy. tsv file. This file and the feature-table. biom file were used in a Perl script designed to summarize the taxonomic information into feature-table-with-taxonomy.txt. Heatmaps were created in R using the pheatmap package and taxa-normalize-pct-per-region.txt file.

Contaminant Filtering

Contaminant sequences were filtered out from the ATCC sample data. Any taxa that were detected in only one of the five technical replicates, detected at less than 0.1% abundance, or both, was considered a contaminant. Filtering was performed on the feature table that was created after open reference OTU clustering using QIIME taxa filter-table. Contaminants are listed in **Supplementary Table S1**.

Generalized Linear Modeling

We used the generalized linear model function in Base R to evaluate statistical differences in alpha diversity and individual taxonomic abundance between fresh versus frozen samples in the clinical cohort. The GLM per feature took the following structure: *log10(feature) ~ fresh/frozen status + specimen ID + hypervariable region*. Regions V8 and V9 were excluded from GLM analysis, and Region V2 was used as the null factor level. The fresh/frozen status of samples was compared, with fresh as baseline factor level set as zero and frozen set as one. The input of "feature" was either an alpha diversity value (Shannon, evenness, observed OTUs or Faith's phylogenetic diversity), or taxonomic abundance of a feature at a specific taxonomic level. Input feature values were log transformed in order to increase stability of values from

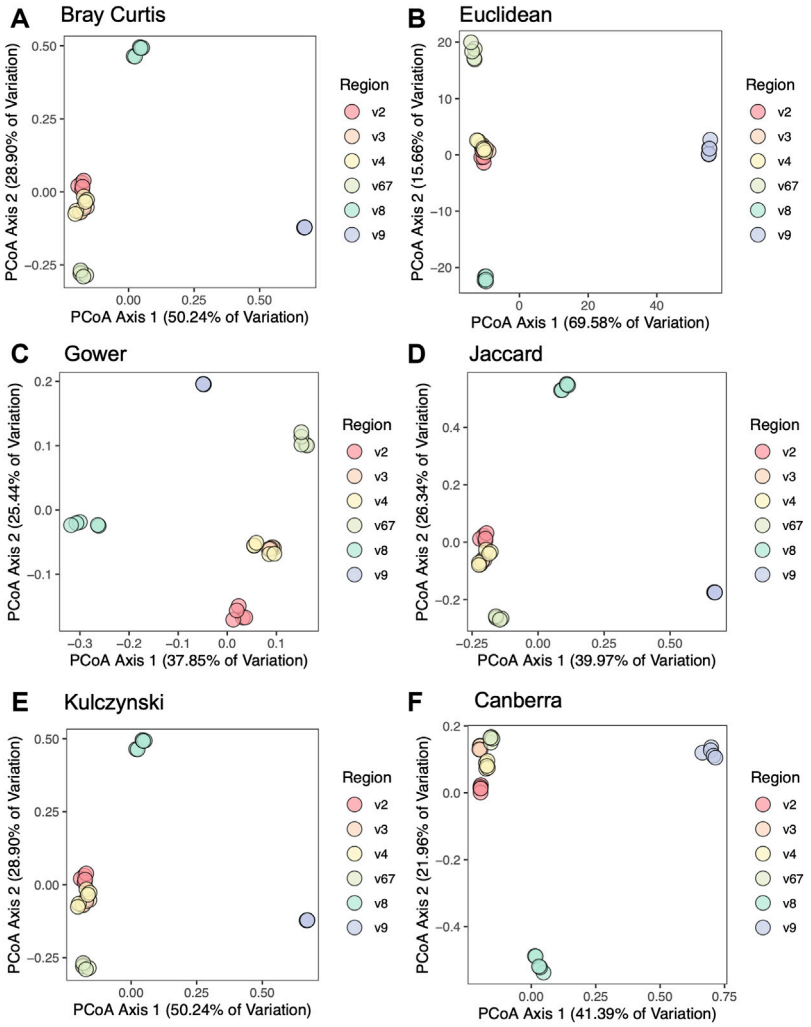


FIGURE 3 | Principal coordinates analysis of mock community samples. PCoA plots are based on distance matrices for **(A)** Bray-Curtis, **(B)** Euclidean, **(C)** Gower, **(D)** Jaccard, **(E)** Kulczynski, and **(F)** Canberra.

person to person when performing statistics. The GLM p-value was obtained by comparing the GLM factor level coefficient to the null hypothesis of zero, which was done *via* a Wald Test.

Data and Code Availability

All sequence files are available in the NCBI Sequence Read Archive (SRA) under Bioproject ID PRJNA738491 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA738491). All codes are available on the public GitHub repository it-workflow (http://github.com/Sfanos-Lab-Microbiome-Projects/it-workflow/).

RESULTS

Mock Community

In order to test our analysis pipeline (**Figure 1**) we prepared libraries and sequenced DNA from a mock microbial community (**Table 1**). A total of five independent replicates from four library preparations of the mock community were sequenced over three sequencing runs. We filtered out low-level contaminants (**Supplementary Table S1**) prior to performing community alpha and beta diversity and taxonomic abundance analyses (see Methods).

We analyzed four different alpha diversity metrics: two measures of evenness (evenness and Shannon diversity), and two measures of richness (Faith's phylogenetic diversity and observed-OTUs) (**Figure 2**). V9 had significantly decreased alpha diversity compared to all regions across all metrics (**Supplementary File S2**). V8 also had significantly decreased Shannon diversity, evenness, and Faith's phylogenetic diversity compared to other regions excluding V9, with two exceptions being that Evenness was not significantly decreased in V8 compared to that of V6-7 and Faith's PD is not significantly decreased in V8 compared to V4 (**Supplementary File S2**).

To compare beta diversity between hypervariable regions and circumvent the issue that OTUs would be region-specific, we used

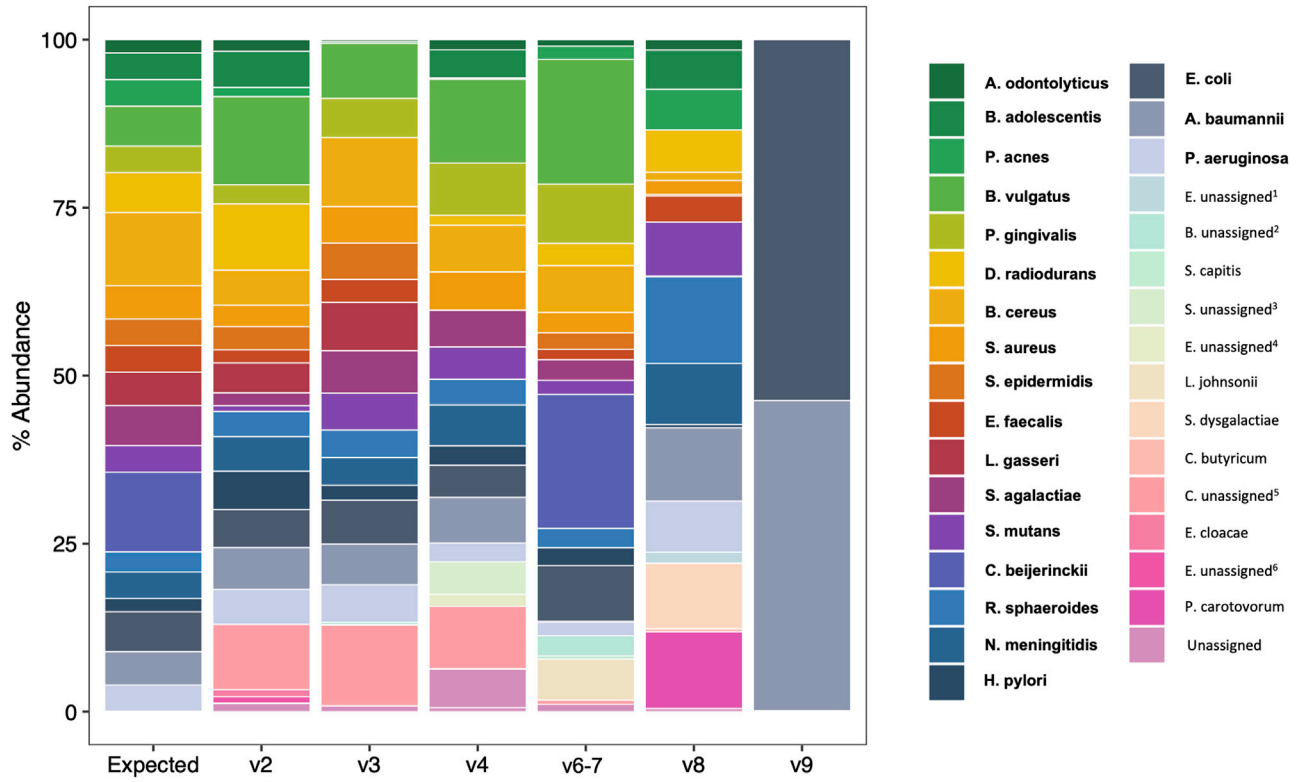


FIGURE 4 | Species-level taxonomic barplots of ATCC 20 Strain Mix sequencing results by 16S rRNA hypervariable region. Bolded taxa are those present in the 20 strain mix. *Enterobacteriaceae* unassigned[1], *Bifidobacterium* unassigned[2], *Staphylococcus* unassigned[3], *Enterococcus* unassigned[4], *Clostridium sensu stricto 1* unassigned[5], *Enterobacter* unassigned[6].

TABLE 2 | Observed species rRNA gene abundance denoted as percent of total.

Species	Expected	V2	V3	V4	V6-7	V8	V9
Acinetobacter baumannii	5.26	6.23	6.06	6.82	0.11	10.90	46.15
Actinomyces odontolyticus	1.75	1.75	0.27	1.51	0.97	1.54	0.00
Bacillus cereus	10.52	5.20	10.29	6.97	6.99	1.21	0.00
Bacteroides vulgatus	6.14	13.11	8.19	12.48	18.57	0.00	0.00
Bifidobacterium adolescentis	4.39	5.37	0.00	4.22	0.00	5.85	0.00
Clostridium beijerinckii	12.28	0.00	0.00	0.00	19.94	0.12	0.00
Deinococcus radiodurans	6.14	9.88	0.00	1.47	3.29	6.31	0.00
Enterococcus faecalis	3.51	1.97	3.41	0.00	1.53	3.88	0.00
Escherichia coli	6.14	5.64	6.52	4.77	8.29	0.00	53.73
Helicobacter pylori	1.75	5.70	2.24	2.90	2.68	0.51	0.00
Lactobacillus gasseri	5.26	4.45	7.21	0.00	0.00	0.00	0.00
Neisseria meningitidis	3.51	5.14	4.13	6.05	0.00	9.07	0.00
Porphyromonas gingivalis	3.51	2.85	5.83	7.77	8.83	0.00	0.00
Propionibacterium acnes	3.51	1.35	0.27	0.17	1.97	6.04	0.00
Pseudomonas aeruginosa	3.51	5.20	5.58	2.79	2.00	7.62	0.00
Rhodobacter sphaeroides	2.63	3.75	4.10	3.83	2.87	12.88	0.00
Staphylococcus aureus	5.26	3.19	5.44	5.71	3.00	2.08	0.00
Staphylococcus epidermidis	4.39	3.44	5.40	0.00	2.49	0.23	0.00
Streptococcus agalactiae	6.14	1.89	6.31	5.46	3.06	0.00	0.00
Streptococcus mutans	4.39	0.87	5.47	4.81	2.09	8.03	0.00
Total Species Identified	20	19	17	16	17	15	2

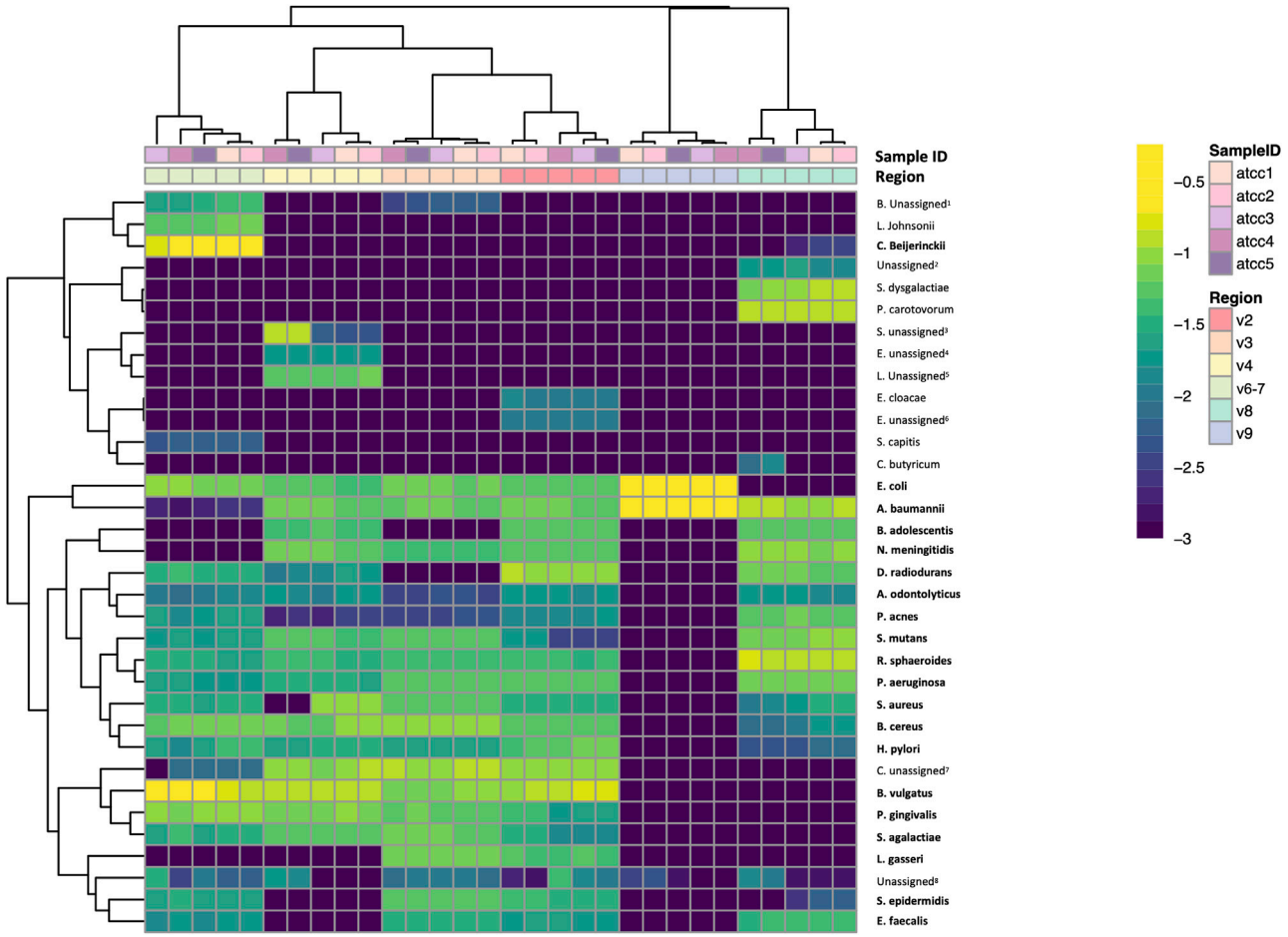


FIGURE 5 | Species-level clustered heatmap of ATCC 20 Strain Even Mix. Bolded taxa are those present in the 20 strain mix. *Bifidobacterium* unassigned[1], *Enterobacteriaceae* unassigned[2], *Staphylococcus* unassigned[3], *Enterococcus* unassigned[4], *Lactobacillus* unassigned[5], *Enterobacter* unassigned[6], *Clostridium sensu stricto 1* unassigned[7], Unassigned at every taxonomic level[8].

taxonomic results from each hypervariable region to create aggregated distance matrices. We assessed six different beta diversity metrics: Canberra, Bray-Curtis, Jaccard, Euclidean, Gower, and Kulczynski. **Figure 3** shows PCoA plots based on six different beta diversity metrics: Bray-Curtis (**Figure 3A**), Euclidean (**Figure 3B**), Gower (**Figure 3C**), Jaccard (**Figure 3D**), Kulczynski (**Figure 3E**), and Canberra (**Figure 3F**). In the plot based on the Canberra distance matrix (**Figure 3F**), the V2, V3, V4, and V6-7 hypervariable regions clustered together, whereas V8 and V9 were distantly separated. This pattern was also observed by the other beta diversity metrics, with V6-7 sometimes also segregating slightly from V2, V3, and V4 which were largely clustered together.

In addition to biodiversity measurements and beta diversity metrics, the percent abundance of the identified organisms after taxonomic classification was evaluated and is given in **Supplementary File S3**. The majority of species were identified by taxonomic classification of the sequences covering each hypervariable region, with the exception of V9 that only positively identified *Escherichia coli* and *Acinetobacter baumannii*. *Clostridium beijerinckii* was the most difficult organism to speciate and was only correctly classified in V6-7 amplicons. The results with hypervariable regions V2, V3, and V4 only identified *Clostridium beijerinckii* at the genus level, V8 misclassified it as *Clostridium butyricum*, and V9 did not identify any Clostridial organisms (**Supplementary File S3**). Aside from *C. beijerinckii*, species misclassification varied by hypervariable region.

We next compared observed versus expected percent abundance by hypervariable region. There are 114 copies of the 16 S rRNA gene in the bacterial genomes comprising the mock community. Therefore, the expected abundance of a given species' rRNA gene is the number of copies in its genome (**Table 1**), divided by 114. Taxonomic bar plots demonstrate the percent abundance of each taxon by hypervariable region compared to expected (**Figure 4**). V2 most closely approximated the overall distribution of species compared to expected and correctly assigned the most species from the mock community (19/20). V3 (17/20), V6-7 (17/20),

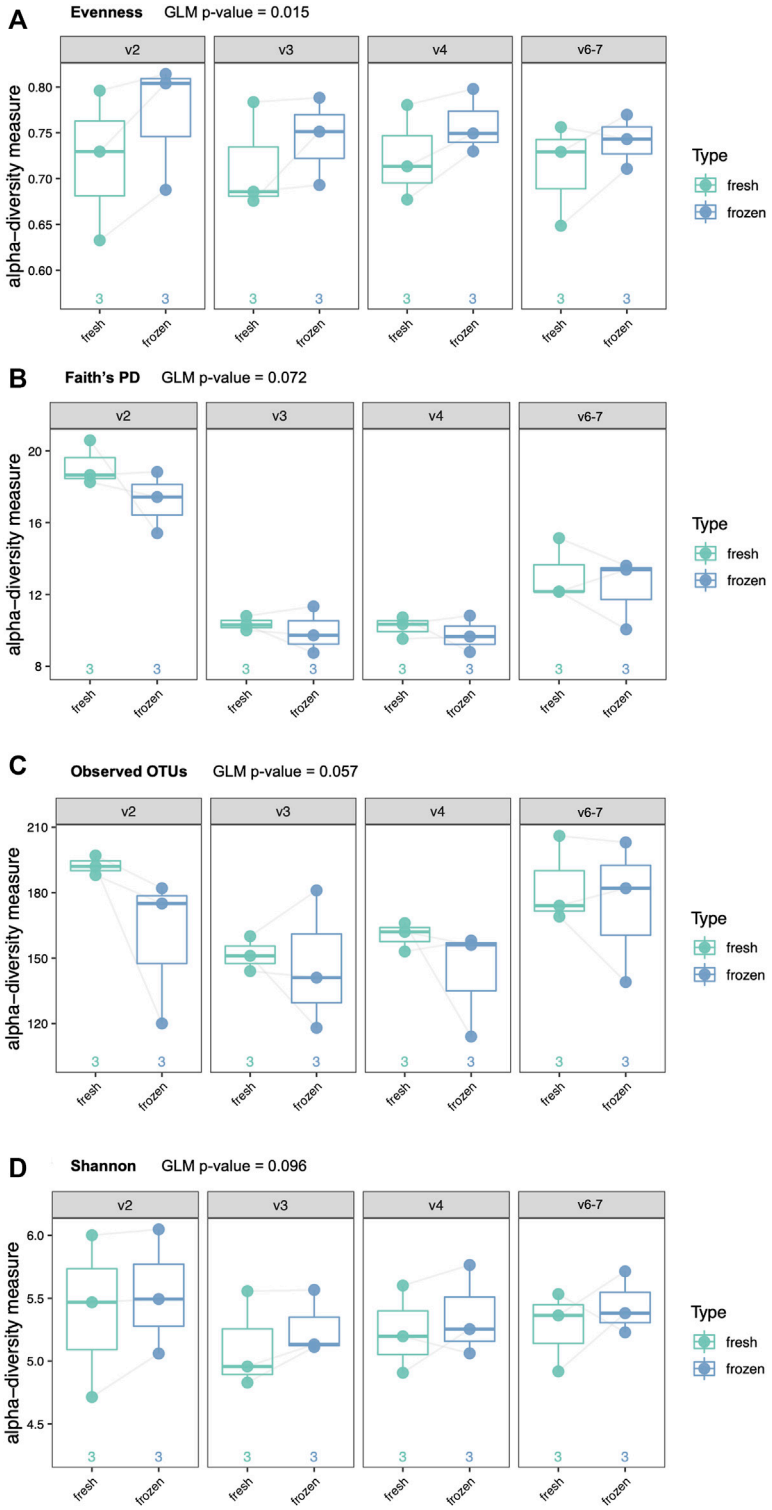


FIGURE 6 | Alpha diversity analyses of six clinical samples by type (fresh or frozen) and hypervariable region. Each patient provided two swabs, one of which was frozen prior to DNA extraction. **(A)** Evenness ($p = 0.015$), **(B)** Faith's phylogenetic diversity ($p = 0.072$), **(C)** Observed Operational Taxonomic Units (OTUs) ($p = 0.067$), **(D)** Shannon diversity ($p = 0.096$).

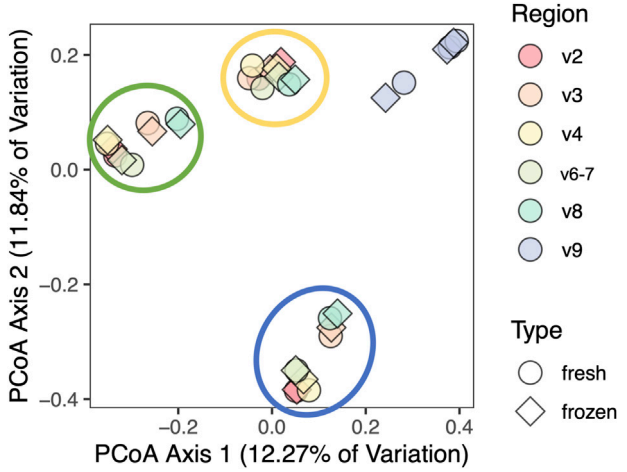


FIGURE 7 | Principal coordinates analysis of clinical cohort using Canberra distance matrix. Samples and regions from the same person are circled, excluding V9. Results cluster by individual and by V9 region (not circled), but not by fresh versus frozen status.

and V4 (16/20) followed closely behind, whereas V8 assigned 15/20, and V9 was only able to identify two species (2/20) (**Table 2**).

Lastly, we performed a clustered heatmap analysis at the species level. The resulting heatmap demonstrated that technical replicates of the mock community sequences cluster by hypervariable region (**Figure 5**). The heatmap visually emphasizes the difference in taxonomic identification in V8 and particularly V9 compared to the other regions. It also highlights misclassifications and which regions were only able to classify taxa to the genus level. Interestingly, the heatmap highlights a few misclassifications or false negatives that occurred in only a subset of the replicates. For example, *Staphylococcus aureus* was classified as *Staphyloccoccus* unassigned in replicates four and five. The OTU tables for these samples indicate that the sequence was truncated prematurely in replicates four and five, indicating the differences in classification here arise from library preparation or sequencing errors rather than downstream data analysis.

Taxonomic Classification of Human Gut Microbiome Culture Collection

Since there appeared to be differing abilities of classification of bacterial species by hypervariable region in our ATCC data set, we next determined if this was the case for a larger pool of bacteria. We plotted out the taxonomic classification results from our *in silico* database validation to visualize whether sensitivity and specificity was region specific (**Supplementary Figure S1**). The sensitivity and mis-classification rates varied with respect to particular species and hypervariable regions. For example, *Bifidobacterium longum* is 100% misassigned when using sequences from V4, but no other region. This region likewise has 0% sensitivity for *B. longum*. Alternatively, *Bifidobacterium bifidum* has high specificity across all hypervariable regions, implying that sensitivity and specificity of taxonomic classification may be increased by using data from multiple hypervariable regions.

Clinical Samples

We next sequenced and analyzed a set of six patient samples in order to demonstrate the use of a generalized linear model (GLM) in an illustrative clinical sample set, incorporating information from multiple hypervariable regions. Hypervariable regions V2, V3, V4, and V6-7 were included in the GLM, while data from the V8 and V9 regions were excluded due to their demonstrated poor performance in identifying species in the mock community (**Figures 2–5**). Samples consisted of duplicate rectal swabs from three participants. DNA was extracted immediately after collection from one rectal swab sample chosen at random from each patient (fresh) and the other sample was frozen at -80°C prior to DNA isolation (frozen). Libraries were prepared in tandem, and all samples were sequenced on the same sequencing run. Sequencing results were processed as outlined above (**Figure 1**).

We performed the same four alpha diversity metrics for the clinical cohort as for the mock community samples (evenness, Shannon diversity, observed OTUs, and Faith's phylogenetic diversity). There were no significant differences in alpha diversity between fresh and frozen samples by Shannon diversity, Faith's phylogenetic diversity or observed OTUs when using a GLM (**Figure 6**). Evenness was slightly increased in frozen samples across all hypervariable regions (adjusted GLM $p = 0.015$).

We aggregated taxonomic results and used them to create Bray-Curtis, Jaccard, Canberra, Euclidean, Gower, and

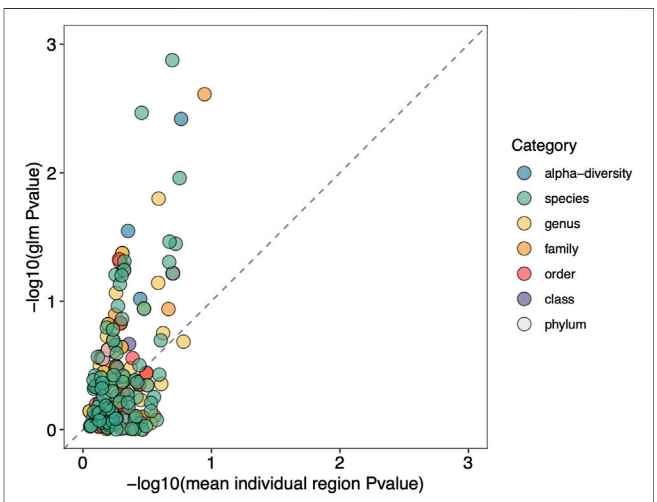


FIGURE 8 | Using a GLM shows enrichment of taxonomic classification sensitivity. GLM p-values for specific taxa are plotted on the *y*-axis, and the mean p-value across all hypervariable regions for the same taxa are plotted on the *x*-axis. *p*-values are log-transformed and multiplied by -1 so that more significant p-values are higher in value. The dashed line indicates where the p-values resulting from the GLM and from individual regions are equal. Enrichment above the dashed line indicates the GLM approach is more sensitive compared to analyzing individual regions.

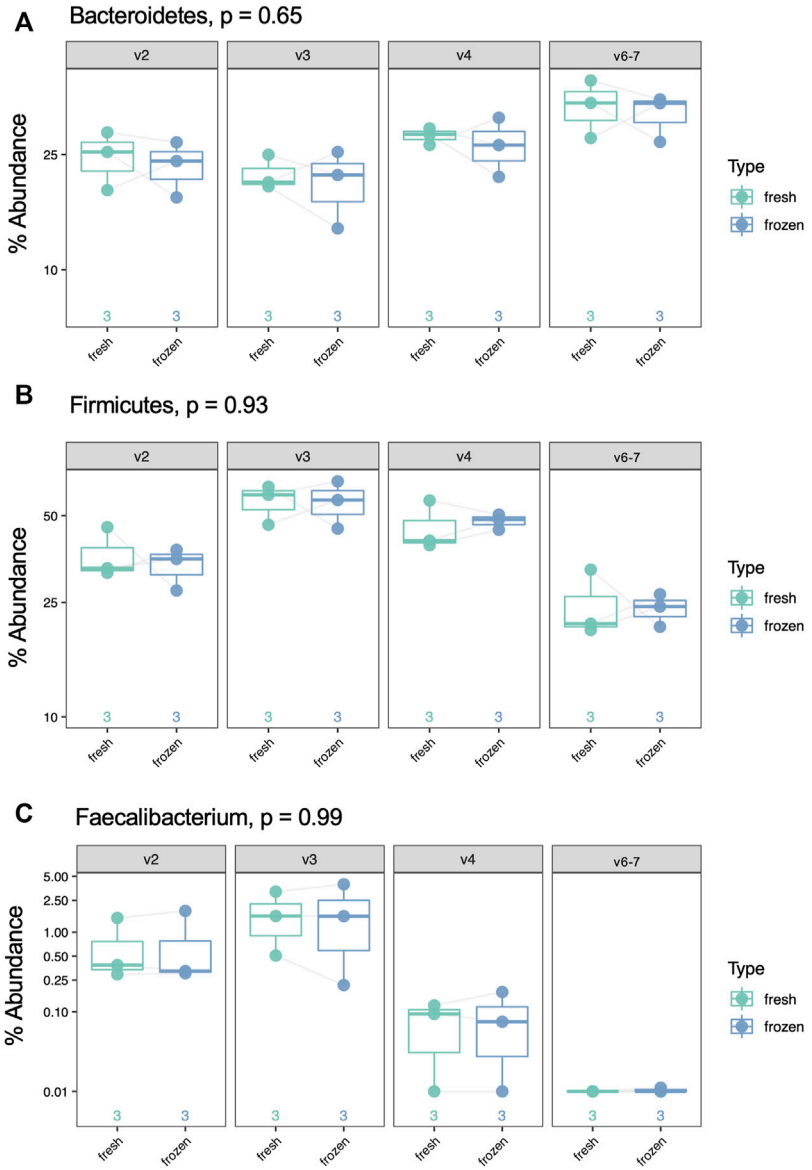


FIGURE 9 | Percent abundance of *Bacteroidetes*, *Firmicutes*, and *Faecalibacterium* by sample type (fresh *vs* frozen) and hypervariable region. *p*-value was calculated with a log-transformed GLM and is false discovery rate-adjusted. **(A)** Bacteriodetes, $p = 0.65$, **(B)** Firmicutes, $p = 0.93$, **(C)** Faecalibacterium, $p = 0.99$.

Kulczynski distance matrices in order to perform combined beta diversity analysis across all hypervariable regions. As demonstrated by the Canberra PCoA plot in **Figure 7**, most variation in beta diversity was due to different individuals and V9 sequences. PERMANOVA analysis of results from each individual hypervariable region demonstrated that total composition does not differ by fresh versus frozen status after adjusting for individual person and region-to-region variation (**Supplementary File S4**).

We next show that using a GLM that incorporates information from multiple variable regions increases the ability to detect significant differences between groups. This is demonstrated in **Figure 8**, where we plot the average p-value for each specific taxon across all hypervariable regions against the p-value obtained for the same taxon when using a GLM. Due to small sample size, we opted to use unadjusted p-values. There is an enrichment of significant p-values when using the GLM as seen by the shift upwards above the dashed line, indicating an increase in sensitivity compared to analyzing individual hypervariable regions.

Using our GLM, we systematically compared abundance of taxa between fresh and frozen samples at multiple levels (phylum, class, order, family, genus, species). As an example, we chose to examine levels of *Firmicutes*, *Bacteroidetes*, and *Faecalibacterium* due to previous reports of differential abundance in fresh verses frozen samples (Bahl et al., 2012; Fouhy et al., 2015). Our results showed no significant differences between these taxa (**Figure 9**)

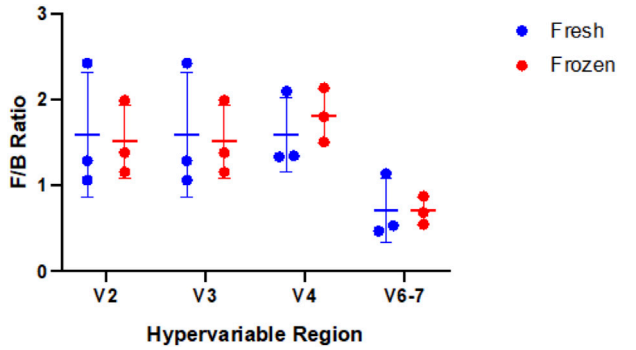


FIGURE 10 | Comparison of *Firmicutes* to *Bacteroidetes* (F/B) ratio in fresh versus frozen samples by hypervariable region. No significant difference was observed between fresh and frozen samples for the hypervariable regions (V2 p-value = 0.87, V3 p-value = 0.87, V4 p-value = 0.51, V6-7 p-value = 0.97).

or *Firmicutes* to *Bacteroidetes* ratios (**Figure 10**). While no concrete conclusions can be made from this data due to small sample size, we demonstrate the utility of the GLM using clinical samples.

DISCUSSION

16 S rRNA sequencing is cost effective, requires relatively low DNA input, and has a number of highly curated reference databases and open-source analysis platforms, making it a common tool for microbiome researchers. PCR amplification using primers that target conserved regions of the 16 S rRNA gene and amplify across hypervariable regions allows amplification of DNA across a widespread taxonomic spectrum and provides unique sequences that can be used for taxonomic classification at higher levels (e.g., family, genus, and species level). Next generation sequencing strategies are often limited to sequencing across only one or at most two of the nine hypervariable regions. The Ion 16 S™ Metagenomics Kit provides the opportunity to prepare libraries containing sequences from seven of the nine hypervariable regions (V2, V3, V4, V6-7, V8, and V9). However, the Ion Reporter analysis pipeline available to Ion 16 S™ Metagenomics Kit users does not allow users to incorporate their own study metadata into analyses and does not allow users to export usable data for downstream analyses, necessitating the development of open-resource analysis tools for data produced from the Ion 16 S™ Metagenomics Kit.

Herein, we report results from sequencing a mock microbial community using the Ion 16 S™ Metagenomics Kit and comparing results from different hypervariable regions. Using a cohort of clinical samples, we demonstrate that taxonomic classification is enhanced by using a generalized linear multivariate model (GLM) that incorporates sequencing data from multiple hypervariable regions.

We first prepared and sequenced five technical replicates of DNA from a twenty strain mock microbial community, and then assessed alpha diversity (evenness, Shannon diversity, observed OTUs, and Faith's phylogenetic diversity) among different hypervariable regions. Even with our limited mock community dataset, we observed hypervariable region-based differences in alpha diversity. Most notably, taxa identified with V9 primers had significantly decreased alpha diversity compared to all other regions across all metrics. V8 results likewise had significantly decreased Shannon Diversity and Faith's PD, suggesting that V8 and V9 are falsely underrepresenting the diversity of the samples.

We performed six different beta diversity metrics (Bray-Curtis, Jaccard, Canberra, Euclidean, Gower, and Kulczynski) to evaluate differences between hypervariable regions. Distance matrices used in beta diversity analyses are generated from OTU tables, however the OTUs identified were not consistent among hypervariable regions. Therefore, in order to compare results between hypervariable regions, we assembled distance matrices using taxonomic results. PCoA analyses demonstrated clustering primarily by hypervariable regions V2, V3, V4, and V6-7. Hypervariable regions V8 and V9 clustered separately from the other regions, again demonstrating the poor performance of amplicon sequencing of these regions in assessing the constituents of the mock community sample.

Consistent with previous reports (Claesson et al., 2010; Cai et al., 2013; Tremblay et al., 2015; Barb et al., 2016), we found that the taxonomic classification results from the mock community samples varied by hypervariable region. Primers targeting the V2, V3, and V6-7 regions identified nearly all the species present in the mock community (19/20, 17/20, and 17/20 respectively), V4 identified 16/20 species, V8 identified 15/20 species, and V9 identified only two (2/20) (**Figure 3**; **Table 2**). Generally, those regions which identified more species present in the mock community also had more evenly distributed observed taxa (i.e., there were no extreme over- or underestimated taxa which skewed the remaining percent abundances, such as in the case of V9).

Errors and biases that contribute to artifacts in PCR-based microbiome studies include sequence artifacts (formation of chimeras or heteroduplexes, or polymerase errors), PCR bias (differing amplification efficiencies of different templates), or biases in the analysis pipeline (poorly discriminatory sequences) (Acinas et al., 2005). Of all OTUs assigned to the V9 region, only two OTUs made up 99.78% of total V9 reads. Therefore, we deduce that the lack of diversity in the region is likely most related to PCR bias. Since V9 lacks sensitivity for many species, we opted to leave this region out of the generalized linear model we used on the clinical samples. V8 also tended to be less sensitive compared to V2, V3, V4, and V6-7, and contributed to variation in the data according to PCoA plots. Therefore, V8 was excluded from further analyses as well. Notably, primer sequences for this kit are not available, and having access to primer sequences in this instance would aid in delving further into why V8 and V9 provided so little information. For others attempting to incorporate a GLM into their analysis, we would recommend against using data from V8 and V9. One must also take into account whether specific regions have increased or decreased sensitivity for specific taxa

of interest when considering which regions to include in your GLM.

Researchers can circumvent the issue of choosing only one hypervariable region to analyze by sequencing multiple hypervariable regions in tandem. Since the sensitivity of each hypervariable region for identifying bacterial taxa varies, combining the results from multiple hypervariable regions for analyses may be misleading. Fuks et al. developed Short MUltiple Regions Framework (SMURF), which combines sequences from multiple PCR amplicons in order to provide one overall set of taxonomic profiling results (Fuks et al., 2018). However, this method is computationally intensive and requires proprietary software. Therefore, to utilize information from multiple hypervariable regions at once and to strengthen confidence in the taxonomic abundance results, we incorporated a generalized linear model (GLM) into alpha diversity and taxonomic abundance analyses.

We demonstrated use of the GLM *via* analysis of a clinical cohort, where each participant donated two rectal swab samples, one of which was processed fresh and the other one frozen prior to DNA extraction. Alpha diversity analysis revealed increased evenness in frozen samples compared to fresh samples. This trend was visualized in results from each individual hypervariable region and was strengthened in the GLM. There was no difference in Shannon's diversity, observed OTUs, and Faith's phylogenetic diversity between fresh and frozen samples which suggests that freezing samples may not affect the ability to detect taxa, but it might alter the detectable abundance of certain taxa. Beta diversity analysis demonstrated clustering of samples by person irrespective of fresh versus frozen status or hypervariable region, with the exception of V9. PERMANOVA analysis confirmed that most of the variation in composition was due to individuals as opposed to storage type. An important limitation of our beta diversity analysis is that in order to compare results from all hypervariable regions in the same analysis, we had to use taxonomic classification as opposed to OTUs. This limits our beta diversity analysis to using only those reads that were assigned taxonomy.

By utilizing a GLM with sequences from our clinical samples, sensitivity to changes between groups was enriched compared to using only one hypervariable region. *p*-values for specific differences in taxa between fresh and frozen samples became significant when utilizing sequences from multiple hypervariable regions, while one region was not powerful enough to detect these differences as observed in **Figure 8**.

Finally, based on the findings above, we compared taxonomic abundance at multiple levels between fresh and frozen samples using a GLM. We found no taxa at any level had significantly different abundance. This is unsurprising based on our small sample size, the fact that alpha and beta diversity were minimally different between sample type, and the fact that other studies show limited differences between fresh verses frozen samples (Bahl et al., 2012; Fouhy et al., 2015). However, *Faecalibacterium* results highlight the important point that not all regions are able to identify a taxon of interest: V6-7 fails to map any reads to this taxon despite its presence in the sample. Thus, even though the true composition of a clinical sample may be unknown, examining redundant data from multiple hypervariable regions may help elucidate the true microbial makeup of the sample, with the caveat that none of the hypervariable regions included vary too significantly from the others to prevent skewing the data.

In conclusion, we propose a method to overcome the issues of analyzing multiple amplicons covering multiple hypervariable regions at once. While this protocol is tailored towards analyzing data generated from the Ion Torrent platform, the approach of sequencing multiple hypervariable regions and analyzing data in parallel could be applied towards Illumina sequencing data, as well. As more tools to analyze more of the 16 S rRNA gene at once become available, it is critical for the microbiome bioinformatics community to come to a consensus as to the proper way to analyze this type of data in order to maintain data quality, and to be able to compare results across different publications.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA738491 http://github.com/Sfanos-Lab-Microbiome-Projects/it-workflow/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Johns Hopkins Medicine Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CJ, LP, and KS contributed to conception and design of the study. SE assisted in sample collection and data curation. CJ, LP, and JW contributed to formal analysis, methodology and data curation. CJ and LP wrote the first draft of the manuscript. CJ, LP, JW, and KS contributed to manuscript revision. All authors approved the submitted version. CJ and LP contributed equally to this work.

FUNDING

This work was supported by The Assistant Secretary of Defense for Health Affairs Endorsed by the Department of Defense through the Prostate Cancer Research Program–Early Investigator Research Award under Award No. W81XWH-18-1-0545 (LP, https://cdmrp.army.mil/pcrp/default) and Prostate Cancer Challenge Award from the Prostate Cancer Foundation under Award No. 16CHAL13 (KS, https://www.pcf.org/science-impact/the-work-we-fund/challenge-awards/). Opinions, interpretations,

conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

This manuscript originally appeared as a preprint on bioRxiv (Jones et al., 2021).

ACKNOWLEDGMENTS

We would like to thank and acknowledge Dr. Angélica Cruz-Lebrón for careful review of the manuscript and helpful suggestions. We also thank the QIIME2 forum community for their help and discussions regarding analyzing multiple hypervariable regions of Ion Torrent data, especially Evan Bolyen, Nicholas Bokulich, Matthew Dillon, Justine Debelius, Colin Brislawn, Jennifer Barb, and Katherine Maki. We would like to thank Jennifer Meyers, Hai Xu, and Kornel Schuebel from the JHU SKCCC Experimental and Computational Genomics Core for their help in generating the sequencing data. Finally, we thank Bradley Toms and Leonardo Varuzza from ThermoFisher for their assistance with library preparation and data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.799615/full#supplementary-material

Supplementary Figure S1 | Taxonomic sensitivity and mis-classification rates of human gut microbiome culture collection by hypervariable region.

Supplementary Table S1 | List of contaminants. **Supplementary File S1**. *In silico* taxonomic validation results.

Supplementary File S2 | Mock community alpha diversity statistics.

Supplementary File S3 | Mock community filtered percent abundance.

Supplementary File S4 | PERMANOVA analysis of fresh *vs* frozen clinical samples.

REFERENCES

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M. F. (2005). PCR-induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Appl. Environ. Microbiol.* 71, 8966–8969. doi:10.1128/aem.71.12.8966-8969.2005

Bahl, M. I., Bergström, A., and Licht, T. R. (2012). Freezing Fecal Samples Prior to DNA Extraction Affects the Firmicutes to Bacteroidetes Ratio Determined by Downstream Quantitative PCR Analysis. *FEMS Microbiol. Lett.* 329, 193–197. doi:10.1111/j.1574-6968.2012.02523.x

Barb, J. J., Oler, A. J., Kim, H.-S., Chalmers, N., Wallen, G. R., Cashion, A., et al. (2016). Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PLoS One* 11, e0148047. doi:10.1371/journal.pone.0148047

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A., and Caporaso, J. G. (2018). q2-sample-classifier: Machine-Learning Tools for Microbiome Classification and Regression. *J. Open Res. Softw.* 3. doi:10.21105/joss.00934

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9

Cai, L., Ye, L., Tong, A. H. Y., Lok, S., and Zhang, T. (2013). Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PLoS One* 8, e53649. doi:10.1371/journal.pone.0053649

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, 581–583. doi:10.1038/nmeth.3869

Claesson, M. J., Wang, Q., O'sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., et al. (2010). Comparison of Two Next-Generation Sequencing Technologies for Resolving Highly Complex Microbiota Composition Using Tandem Variable 16S rRNA Gene Regions. *Nucleic Acids Res.* 38–e200. doi:10.1093/nar/gkq873

Clemmons, B. A., Voy, B. H., and Myer, P. R. (2019). Altering the Gut Microbiome of Cattle: Considerations of Host-Microbiome Interactions for Persistent Microbiome Manipulation. *Microb. Ecol.* 77, 523–536. doi:10.1007/s00248-018-1234-9

Debelius, J. W., Robeson, M., Hugerth, L. W., Boulund, F., Ye, W., and Engstrand, L. (2021). A Comparison of Approaches to Scaffolding Multiple Regions along the 16S rRNA Gene for Improved Resolution. bioRxiv, 2021.2003.2023.436606.

Faith, D. P., Minchin, P. R., and Belbin, L. (1987). Compositional Dissimilarity as a Robust Measure of Ecological Distance. *Vegetatio* 69, 57–68. doi:10.1007/bf00038687

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., et al. (2019). A Human Gut Bacterial Genome and Culture Collection for Improved Metagenomic Analyses. *Nat. Biotechnol.* 37, 186–192. doi:10.1038/s41587-018-0009-7

Fouhy, F., Deane, J., Rea, M. C., O'Sullivan, Ó., Ross, R. P., O'Callaghan, G., et al. (2015). The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. *PLoS One* 10, e0119355. doi:10.1371/journal.pone.0119355

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA Gene Variable Regions Enables High-Resolution Microbial Community Profiling. *Microbiome* 6, 17. doi:10.1186/s40168-017-0396-x

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* 40, 237–264. doi:10.1093/biomet/40.3-4.237

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44, 223–270.

Jones, C. B., White, J. R., Ernst, S. E., Sfanos, K. S., and Peiffer, L. B. (2021). Incorporation of Data from Multiple Hypervariable Regions when Analyzing Bacterial 16S rRNA Sequencing Data. bioRxiv.

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors that Structure Microbial Communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi:10.1128/aem.01996-06

Lozupone, C., and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi:10.1128/aem.71.12.8228-8235.2005

Myer, P. R., Mcdaneld, T. G., Kuehn, L. A., Dedonder, K. D., Apley, M. D., Capik, S. F., et al. (2020). Classification of 16S rRNA Reads Is Improved Using a Niche-specific Database Constructed by Near-Full Length Sequencing. *PLoS One* 15, e0235498. doi:10.1371/journal.pone.0235498

Pinto, A. J., and Raskin, L. (2012). PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS One* 7, e43093. doi:10.1371/journal.pone.0043093

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5, e9490. doi:10.1371/journal.pone.0009490

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun Metagenomics, from Sampling to Analysis. *Nat. Biotechnol.* 35, 833–844. doi:10.1038/nbt.3935

Ranjan, R., Rani, A., Metwally, A., Mcgee, H. S., and Perkins, D. L. (2016). Analysis of the Microbiome: Advantages of Whole Genome Shotgun versus 16S Amplicon Sequencing. *Biochem. Biophysical Res. Commun.* 469, 967–977. doi:10.1016/j.bbrc.2015.12.083

Ritari, J., Salojärvi, J., Lahti, L., and De Vos, W. M. (2015). Improved Taxonomic Assignment of Human Intestinal 16S rRNA Sequences by a Dedicated Reference Database. *BMC Genomics* 16, 1056. doi:10.1186/s12864-015-2265-y

Sanschagrin, S., and Yergeau, E. (2014). Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons. *J. Vis. Exp.* doi:10.3791/51709

Shannon, C. E. (2001). A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.* 5, 3–55. doi:10.1145/584091.584093

Shrestha, E., White, J. R., Yu, S.-H., Kulac, I., Ertunc, O., De Marzo, A. M., et al. (2018). Profiling the Urinary Microbiome in Men with Positive versus Negative Biopsies for Prostate Cancer. *J. Urol.* 199, 161–171. doi:10.1016/j.juro.2017.08.001

Sorensen, T. A. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and its Application to Analyses of the Vegetation on Danish Commons. *Biol. Skar.* 5, 1–34.

Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., et al. (2015). Primer and Platform Effects on 16S rRNA Tag Sequencing. *Front. Microbiol.* 6, 771. doi:10.3389/fmicb.2015.00771

Conflict of Interest: JW has financial and/or other relationship with Resphera Biosciences. There are no patents, products in development or marketed products associated with this research to declare.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Genetics

METHODS
published: 01 April 2022
doi: 10.3389/fgene.2022.841764

MiRKAT-MC: A Distance-Based Microbiome Kernel Association Test With Multi-Categorical Outcomes

Zhiwen Jiang[1], Mengyu He[2], Jun Chen[3], Ni Zhao[4] and Xiang Zhan[5]**

[1]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, United States, [2]Department Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, United States, [3]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States, [4]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States, [5]Department of Biostatistics, School of Public Health and Beijing International Center for Mathematical Research, Peking University, Beijing, China

Increasing evidence has elucidated that the microbiome plays a critical role in many human diseases. Apart from continuous and binary traits that measure the extent or presence of a disease, multi-categorical outcomes including variations/subtypes of a disease or ordinal levels of disease severity are commonly seen in clinical studies. On top of that, studies with clustered design (i.e., family-based and longitudinal studies) are popular alternatives to population-based ones as they are able to identify characteristics on both individual and population levels and to investigate the trajectory of traits of interest over time. However, existing methods for microbiome association analysis are inadequate to handle multi-categorical outcomes, neither independent nor clustered data. We propose a microbiome kernel association test with multi-categorical outcomes (MiRKAT-MC). Our method is versatile to deal with both nominal and ordinal outcomes for independent and clustered data. In addition, it incorporates multiple ecological distances to allow for different association patterns between outcomes and microbiome compositions to be incorporated. A computationally efficient pseudo-permutation strategy is used to evaluate the statistical significance. Comprehensive simulations show that MiRKAT-MC preserves the nominal type I error and increases statistical powers under various scenarios and data types. We also apply MiRKAT-MC to real data sets with nominal and ordinal outcomes to gain biological insights. MiRKAT-MC is easy to implement, and freely available via an R package at https://github.com/Zhiwen-Owen-Jiang/MiRKATMC with a Graphical User Interface through R Shinny also available.

Keywords: beta-diversity, longitudinal studies, microbiome association analysis, multi-categorical outcomes, kernel association test

OPEN ACCESS

Edited by:
Harinder Singh,
J. Craig Venter Institute, United States

Reviewed by:
Ximing Xu,
Children's Hospital of Chongqing Medical University, China
Rajesh Kumar,
National Institutes of Health (NIH), United States

****Correspondence:***
Ni Zhao
nzhao10@jhu.edu
Xiang Zhan
zhanx@bjmu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *22 December 2021*
Accepted: *10 March 2022*
Published: *01 April 2022*

Citation:
Jiang Z, He M, Chen J, Zhao N and Zhan X (2022) MiRKAT-MC: A Distance-Based Microbiome Kernel Association Test With Multi-Categorical Outcomes. Front. Genet. 13:841764. doi: 10.3389/fgene.2022.841764

1 INTRODUCTION

The diverse microbial cells including bacteria, archaea, and fungi that colonize the mucosal and skin environment constitute the human microbiome (Gilbert et al., 2018). It is broadly acknowledged that the human microbiome and its interaction with the immune, endocrine, and nervous systems are associated with a variety of illnesses, ranging from inflammatory bowel disease (Ni et al., 2017), to cancer (Kostic et al., 2013a), and to major depressive disorder (Jiang et al., 2015). A key step in investigating the relationship between microbiome and human disorders lies in quantifying the

taxonomic composition. Currently, the most commonly used method is through the sequencing of the 16S ribosomal RNA gene, which, as a biomarker, is present in all prokaryotic cells and reflects the evolutionary distance between distinct genomes. Computationally, the 16S rRNA sequencing tags can be assigned into Operational Taxonomic Units (OTU) or Amplicon Sequence Variants (ASV) as computational surrogate of microbial taxa (Schloss, 2010; Callahan et al., 2016). Through sequencing, the microbial community can be directly quantified, without the need of labor-intensive bacterial culturing. For instance, the disparity between microbiome communities from two samples can be assessed via an ecological distance/dissimilarity metric, such as the UniFrac distance (Lozupone and Knight, 2005) and the Bray-Curtis dissimilarity (Bray and Curtis, 1957).

Identifying links between microbiome and diseases is often achieved by microbiome-wide association studies (MWAS) (Kostic et al., 2013b), which in turn provide insight into the biological mechanisms of human health and disease conditions. The data type of the investigated outcomes varies from study to study. Typically, samples can be dichotomized as cases and controls when exploring human diseases. For example, (Naseribafrouei et al., 2014) discovered potential correlation between depression and fecal microbiota, where study participants were classified as depression vs. non-depression. On the other hand, multi-categorical (nominal or ordinal) outcomes are also frequently encountered and investigated in many microbiome studies. For instance, Scher et al. (Scher et al., 2013) explored the association between rheumatoid arthritis (RA) and gut microbiota by recruiting patients with three different categories of arthritis: new-onset RA, treated RA, and psoriatic arthritis (PsA). Parikh et al. (Parikh et al., 2020) investigated the association between Apolipoprotein E (APOE) alleles and gut microbiome in murine models, where the APOE gene encodes a major cholesterol carrier protein that supports lipid transport and injury repair in the brain. Polymorphism in APOE gene is a major risk for developing Alzheimer disease. In this study, the APOE gene was coded as a nominal variable of different genotypes (APOE2 APOE3, and APOE4). Furthermore, Schirmer et al. Schirmer et al. (Schirmer et al., 2018) investigated the association between severity of ulcerative colitis and gut microbiome, where disease severity was treated as an ordinal variable with four levels: inactive, mild, moderate and severe.

Association analysis between a host trait and microbiome compositions can be generally addressed by PERMANOVA (Anderson, 2001), which partitions the total variation across the microbiome data cloud in the space of a chosen dissimilarity measure into multiple directions. PERMANOVA is able to accommodate both binary and multi-categorical outcomes, but fails to incorporate multiple distance metrics, where distinct distances capture distinct underlying association patterns and therefore are more powerful under different circumstances. Hence, Tang et al. (Tang et al., 2016) proposed PERMANOVA-S to incorporate multiple distance metrics into a single test. However, it is not adequate to multi-categorical outcomes unless we combine multiple categories into a binary variable, which potentially leads to significant power loss. An alternative to PERMANOVA is the family of microbiome regression-based kernel association tests (MiRKAT) (Zhao et al., 2015; Wilson et al., 2021). Utilizing the classic mixed effect models, the MiRKAT approaches summarize the microbiome structure as a kernel similarity matrix (constructed through the sample-sample distance metric) and model it as a random effect. Adjusting for covariates is straightforward in this framework. The association test is conducted via a variance component score test with p-value calculated in multiple ways, including analytical (Chen et al., 2016; Zhan et al., 2017a), permutation (Koh et al., 2019) and fast pseudo-permutation approaches (Zhan et al., 2017b). However, existing MiRKAT tests are not able to accommodate multi-categorical outcomes.

Beyond population-based studies in which all samples are independent, nowadays, researchers frequently collect microbiome data that are clustered or longitudinal in nature. For instance, Goodrich et al. (Goodrich et al., 2014) collected stool samples from female twins in the United Kingdom to investigate the relationship between obesity and gut microbiome. Flores et al. (Flores et al., 2014) explored the effect of antibiotic use on temporal variability of the microbiome diversity and community structure in gut, palm and tongue. Methods available to address correlated outcomes in microbiome studies burgeoned in the recent years (Chen and Li, 2016; Zhan et al., 2018; Zhang et al., 2018; Koh et al., 2019). For instance, GLMM-MiRKAT (Koh et al., 2019) extends MiRKAT for continuous, binary and count outcomes in longitudinal studies. It adopts kernel regression-based generalized linear mixed models to construct variance component tests and uses permutations to calculate the p-value. Unfortunately, only exchangeable clusters which contain identical number of observations and the same time points can be permuted in this approach. Thus, the permutation procedure will be very complicated and inefficient for unbalanced study designs. On top of that, permutation tends to be computationally intensive when the sample size increases (especially for studies with multi-categorical outcomes) or when small p-values are needed for multiple comparison adjustment. These drawbacks also exist for PERMANOVA.

In this paper, we propose a new distance-based microbiome kernel association test for multi-categorical outcomes (MiRKAT-MC), when samples are independent or clustered. MiRKAT-MC works for both nominal and ordinal outcomes, through the use of the generalized logit model (GLM) and the proportional odds model (POM), respectively. We utilize a fast pseudo-permutation technique (Zhan et al., 2017b) to calculate p-values. This approach features several advantages over its potential competitors: 1) it avoids the complication in designing a suitable permutation scheme for inference; 2) it is computationally efficient and much faster than direct permutations; 3) it controls the type I error and maintains high statistical power compared to the analytical approach. For the last point, due to the small sample size and the over-dispersion in microbiome data, it is quite difficult to approximate

the MiRKAT test statistics, especially for clustered/longitudinal data and for outcomes that are not normally distributed.

Another common challenge in distance-based methods lies in how to select an appropriate ecological distance to construct the kernel, because the statistical power highly depends on a proper kernel to capture the underlying association pattern. Attempting multiple kernels and cherry-picking the smallest p-value yields inflated type I errors. On the other hand, naively adjusting the results by Bonferroni correction will reduce the statistical power substantially, mainly because the individual tests are highly correlated. We propose an omnibus test that combines the individual p-values from tests with different kernels through the harmonic mean procedure (HMP) (Wilson, 2019). The omnibus test is not necessarily the most powerful one: which test is the most powerful depends on the true nature of association, which is unknown prior to analysis. Nevertheless, our omnibus test is robust regardless of the real association pattern in that it loses little power compared to the most powerful one, and is much more powerful than choosing an inappropriate kernel.

In summary, the goal of this paper is to introduce novel statistical methods to examine the association between a multi-categorical outcome (both nominal and ordinal) and microbiome composition under different study designs (e.g., independent design, clustered design). Our major contributions are two-fold. First, we have cast the association analysis between a multi-categorical outcome and microbiome composition into frameworks of generalized logit models and proportional odds models (with additional random effects accounting for within-cluster correlations for clustered design). Our second contribution is proposing a robust p-value calculation procedure via a novel fast pseudo-permutation technique (Zhan et al., 2017b), avoiding the complicated and time-consuming permutation approach yet providing valid statistical inference. Finally, we provide a free R software to implement our proposed methods. It is a useful tool for microbiome researchers to investigate the relationship between the microbiome community and a multi-categorical outcome under a wide range of study designs, which was not readily available before.

2 MATERIALS AND METHODS

To associate microbiome compositions with a multi-categorical outcome, we build upon generalized logit models (GLM) for nominal outcomes and proportional odds models (POM) for ordinal outcomes, and relate the microbiome profile with the outcome through the flexible semi-parametric kernel machine regression framework (Zhao et al., 2015). Our proposed MiRKAT-MC includes MiRKAT-MCN (for nominal outcomes) and MiRKAT-MCO (for ordinal outcomes). For both tests, we propose two versions, one for independent samples and another for clustered/longitudinal samples through the use of additional random effects in the generalized logit mixed model (GLMM) or the proportional odds mixed model (POMM).

2.1 GLM and POM for Independent Data

We first describe the GLM and POM model without considering the high dimensional microbiome data. Let $\boldsymbol{Y}_i$ denote the multi-categorical outcome with total J categories for the i-th subject. Here, bmY_i is a vector with the j-th element being y_{ji}, a binary variable denoting whether the i-th sample belongs to the j-th category, $i = 1, \ldots, N, j = 1, \ldots, J$. That is, $y_{ji} = 1$ means subject i is of category j and otherwise, $y_{ji} = 0$. In practice, y_{ji} can represent any mutually-exclusive categorical traits (nominal and ordinal), such as subtypes of cancers and increasing levels of disease severity that $\sum_{j=1}^{J} y_{ji} = 1$. From a probability perspective, $\boldsymbol{Y}_i$ can be considered as from a multinomial distribution with J categories. Let $\pi_j(\boldsymbol{x}_i) = \Pr(y_{ji} = 1|\boldsymbol{x}_i)$ be the conditional probability that subject i is of category j with $\sum_j \pi_j(\boldsymbol{x}_i) = 1$, where $\boldsymbol{x}_i$ denotes the set of covariates that we want to associate Y_i with (such as race, gender and age). If bmY_i is nominal, we can set the last category J as a reference without loss of generalization, and form the following GLM:

$$\log \frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)} = \alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i, \tag{1}$$

where $j = 1, \ldots, J-1$. The left-hand side of **Eq. 1** is the logit of a conditional probability, and each coordinate of $\boldsymbol{\beta}_j$ represents the increase in log-odds of falling into category j vs. the reference category J resulting from a one-unit increase in the corresponding covariate while holding the other covariates constant. This model simultaneously describes the effects of $\boldsymbol{x}_i$ on all outcome categories in contrast to the reference. In this model, parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, J-1$ can be different among categories. If the categories are ordinal, we can utilize the order information and form the following POM:

$$\text{logit}\big(\nu_j(\boldsymbol{x}_i)\big) = \log \frac{\nu_j(\boldsymbol{x}_i)}{1 - \nu_j(\boldsymbol{x}_i)} = \alpha_j + \boldsymbol{\beta}' \boldsymbol{x}_i, \tag{2}$$

where $j = 1, \ldots, J-1$, and

$$\nu_j(\boldsymbol{x}_i) = \sum_{h=1}^{j} \Pr\big(y_{hi} = 1|\boldsymbol{x}_i\big) = \pi_1(\boldsymbol{x}_i) + \cdots + \pi_j(\boldsymbol{x}_i).$$

Here, $\nu_j(\boldsymbol{x}_i)$ is the conditional cumulative probability, and the corresponding response, defined by $\tilde{y}_{ji} = \sum_{h=1}^{j} y_{hi}$, is called the cumulative response. The ordinal information is thus utilized in the way that the original categories enter the groups in a sequence. In contrast to GLM, $\boldsymbol{\beta}$ here keeps constant across $J-1$ logits and the intercepts have to satisfy $\alpha_1 < \ldots < \alpha_{J-1}$ in the proportional odds model.

Finally, we notice that there are other recent attempts to develop association analysis for multi-categorical outcomes using multinomial logistic regression (i.e., GLM model (1)), usually in the context of genome wide association studies (He et al., 2021; Liu et al., 2021). Despite the shared motivations, MiRKAT-MC is distinct from existing methods in multiple aspects. First, none of the existing approaches specifically models ordinal outcomes and thus MiKAT-MC under POM is statistically novel. Second, MiRKAT-MC includes options that utilize GLMM and POMM (described **Section 2.2**) to accommodate non-independent data from more complicated

study designs. Last, our pseudo-permutation approach for obtaining p-values is novel and tends to outperform the asymptotic results as in existing methods when sample sizes are small, which is usually the case in microbiome data.

2.2 GLMM and POMM for Clustered/ Longitudinal Data

Similarly, we first describe the GLMM and POMM models without considering the complex microbiome data. Suppose cluster i has m_i observations. Let $\boldsymbol{Y}_{ik} = (y_{1ik}, \ldots, y_{Jik})'$ represent the multi-categorical outcome of the k-th observation in cluster i, $i = 1, \ldots, n$, $k = 1, \ldots, m_i$ and $N = \sum_{i=1}^{n} m_i$ be the total number of observations in the study. Following notations in the previous section, let $\pi_j(\boldsymbol{x}_{ik}|\boldsymbol{b}_{ji}) = \Pr(y_{jik} = 1|\boldsymbol{x}_{ik}, \boldsymbol{b}_{ji})$ and setting the J-th category as reference, the GLMM for clustered/longitudinal data can be written as:

$$\log\frac{\pi_j\left(\boldsymbol{x}_{ik}|\boldsymbol{b}_{ji}\right)}{\pi_J\left(\boldsymbol{x}_{ik}|\boldsymbol{b}_{ji}\right)} = \boldsymbol{\alpha}_j + \boldsymbol{x}_{ik}'\boldsymbol{\beta}_j + \boldsymbol{u}_{ik}'\boldsymbol{b}_{ji}, \tag{3}$$

where $\boldsymbol{x}_{ik} = (x_{ik1}, \ldots, x_{ikq})'$ denote covariates and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jq})'$ are corresponding regression coefficients, $\boldsymbol{u}_{ik}$ is the design matrix for the random effect term $\boldsymbol{b}_{ji}$. We introduce $\boldsymbol{b}_{ji}$ to model correlations among observations within cluster i of category j. The model definition is completed by specifying the distribution of the random effect $\boldsymbol{b}_{ji} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_j)$, where the variance-covariance matrix $\boldsymbol{G}_j$ for the j-th category is unstructured. We also allow $\boldsymbol{b}_{ji}$ to be correlated across categories.

The corresponding POMM for ordinal outcomes is as follows:

$$\text{logit}\left(\nu_j(\boldsymbol{x}_{ik}|\boldsymbol{b}_i)\right) = \boldsymbol{\alpha}_j + \boldsymbol{x}_{ik}'\boldsymbol{\beta} + \boldsymbol{u}_{ik}'\boldsymbol{b}_i. \tag{4}$$

One main difference between models (**Eqs. 3**, **4**) lies in model (**Eq. 4**) restricts $\boldsymbol{b}_i$ to be identical across category comparisions, and thus $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$ with a fixed variance-covariance matrix $\boldsymbol{G}$. Here, we essentially assume that the random effects across the ordered categories are the same, which guarantees in proportional odds. Specifically, for a fixed cluster i, the random effect $\boldsymbol{b}_i$ has identical value across different categories j. But for different clusters i and i', $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i'}$ may be different and both have normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$. The variance-covariance matrix $\boldsymbol{G}$ is unstructured as well. The same constraints for $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}$ as in model (**Eq. 2**) also apply in the POMM model (**Eq. 4**).

2.3 Microbiome Association Analysis Under Models for Multi-Categorical Variables

We extend the previous described models to incorporate the complex microbiome data. For independent data, let $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})'$ be the composition of p OTUs for sample i (subject to appropriate normalization and transformation). We relate the multivariate outcome to the microbiome community and the covariates with the following model

$$\eta_{ji} = \alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}_j + h_j(\boldsymbol{z}_i), \tag{5}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, J$, where $\eta = g(\cdot)$ and $g(\cdot)$ is a link function. For GLM, $g(\pi_{ji}) = \log(\pi_{ji}/\pi_{Ji})$, $\pi_{ji} = E(y_{ji}|h_{ji})$, and $h_{ji} = h_j(\boldsymbol{z}_i)$; for POM, $g(\nu_{ji}) = \log\{\nu_{ji}/(1 - \nu_{ji})\}$, $\nu_{ji} = E(\tilde{y}_{ji}|h_{ji})$ is the conditional mean of the cumulative response $\tilde{y}_{ji}$. $h_j(\cdot)$ are unknown real functions corresponding to the effects of microbiome on the j-th category. For POM, $h_j(\cdot)$ are identical across categories, and α_j and $\boldsymbol{\beta}_j$ are subject to the constraints described in model (**Eq. 2**).

For clustered studies, let y_{jik} be a binary variable denoting whether the k-th observation of the i-th cluster belongs to the j-th category, where $k = 1, \ldots, m_i$, $i = 1, \ldots, n$ and $j = 1, \ldots, J$. We let $N = \sum_{i=1}^{n} m_i$ be the total number of observations. $\boldsymbol{z}_{ik} = (z_{ik1}, \ldots, z_{ikp})'$ represent p OTUs for the k-th observation in the i-th cluster. The mixed effect model proceeds as

$$\eta_{jik} = \alpha_j + \boldsymbol{x}_{ik}'\boldsymbol{\beta}_j + \boldsymbol{u}_{ik}'\boldsymbol{b}_{ji} + h_j(\boldsymbol{z}_{ik}), \tag{6}$$

where $\eta_{jik} = g[E(y_{jik}|\boldsymbol{b}_{ji}, h_{jik})]$, $h_{jik} = h_j(\boldsymbol{z}_{ik})$, and $g(\cdot)$ is the same link function as model (**Eq. 5**). To illustrate our methodology, we here give some specific examples of the random effects $\boldsymbol{u}_{ik}$. When $u_{ik} = 1$, b_{ji} is the random intercept which can be assumed normally distributed $\sim \mathcal{N}(0, g_{jj})$. When $\boldsymbol{u}_{ik} = (1, t_{ik})'$, where t_{ik} is the time for the k-th observation in the i-th cluster (for longitudinal studies), $\boldsymbol{b}_{ji} = (b_{ji1}, b_{ji2})'$ denote the random intercept and random slope with a bivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_{jj})$, where $\boldsymbol{G}_{jj} = \begin{pmatrix} g_{jj11} & g_{jj12} \\ g_{jj21} & g_{jj22} \end{pmatrix}$. Usually, $\boldsymbol{G}_{jj}$ is specified as "unstructured" in generalized linear mixed effect models, providing much flexibility to capture cluster specific correlations. Again, for POMM, α_j, $\boldsymbol{\beta}_{jm}$, and $\boldsymbol{b}_{ji}$ are subject to the constraints described in model (**Eq. 4**), and $h_{jik}(\cdot)$ should be identical across categories.

Our primary goal is to test the null hypothesis H_0: $h_1(\cdot) = \ldots = h_{J-1}(\cdot) = 0$ in **Eq. 5**, **6**. One feasible approach is to develop such a test leveraging the kernel machine regression-based association analysis framework (Zhao et al., 2015). Through the critical connection between kernel machine regression and mixed models (Liu et al., 2007), $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{J-1})'$ can be considered as random effect with mean $\boldsymbol{0}$ and variance $\boldsymbol{K}^{\star}$. We assume that each $\boldsymbol{h}_j = (h_{j1}, \ldots, h_{jN})'$ for independent data (or $\boldsymbol{h}_j = (h_{j11}, \ldots, h_{j1m_1}, h_{j21}, \ldots, h_{jnm_n})'$ for clustered data) is independent and is of the same (multivariate) distribution. In such a case, $\boldsymbol{K}^{\star} = \boldsymbol{I}_{J-1} \otimes \tau\boldsymbol{K}$, where $\boldsymbol{I}_{J-1}$ denote $(J-1)$-th order identity matrix, τ is an unspecified constant, $\boldsymbol{K}$ is an $N \times N$ kernel matrix, and $\otimes$ denotes Kronecker product. Following (Zhao et al., 2015), the kernel matrix can be easily constructed by a specific ecological distance matrix $\boldsymbol{D}$

$$\boldsymbol{K} = -\frac{1}{2}\left(\boldsymbol{I}_N - \frac{\boldsymbol{1}_N\boldsymbol{1}_N'}{N}\right)\boldsymbol{D}^2\left(\boldsymbol{I}_N - \frac{\boldsymbol{1}_N\boldsymbol{1}_N'}{N}\right), \tag{7}$$

where $\boldsymbol{1}_N$ is a vector of 1's and $\boldsymbol{I}_N$ is the identity matrix.

Typical distance measures for microbiome data include the Bray-Curtis dissimilarity, the weighted, unweighted or generalized UniFrac distances (Lozupone and Knight, 2005). The kernel matrix defined by **Eq. 7** measures sample-pairwise similarities. Using this transformation, ecological information (e.g., taxonomic or the phylogenetic relationship between taxa)

encoded in the distance $\boldsymbol{D}$ is preserved in $\boldsymbol{K}$, and thus in the functions of microbiome effect $h_j(\cdot)$'s (which are assumed to be in the space spanned by $\boldsymbol{K}$). As demonstrated in previous studies, the embedding of such ecological information may boost statistical power for detecting an underlying association under many scenarios (Zhao et al., 2015). Here, we first focus the simpler case in which a single distance (e.g., Bray-Curtis dissimilarity) is considered. Omnibus test utilizing multiple kernels will be described later in this session.

To develop the distance-based kernel association test, we further translate association analysis working model (**Eqs. 5**, **6**) into matrix language. For independent data,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h}, \tag{8}$$

where $\boldsymbol{\eta} = (\eta_{11}, \eta_{12}, \ldots, \eta_{1N}, \ldots, \eta_{J-1,1}, \ldots, \eta_{J-1,N})'$, $\boldsymbol{X} = \boldsymbol{I}_{J-1} \otimes \begin{bmatrix} 1 & \boldsymbol{x}_1' \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_N' \end{bmatrix}$, $\boldsymbol{\beta} = (\alpha_1, \boldsymbol{\beta}_1', \ldots, \alpha_{J-1}, \boldsymbol{\beta}_{J-1}')'$, $\boldsymbol{h} = (h_{11}, h_{12}, \ldots, h_{1N}, \ldots, h_{J-1,1}, \ldots, h_{J-1,N})'$ is distributed as multivariate normal with mean zero and covariance matrix $\boldsymbol{K}^\star = \boldsymbol{I}_{J-1} \otimes \tau\boldsymbol{K}$. Hence, testing H_0: $\boldsymbol{h} = \boldsymbol{0}$ is equivalent to testing H_0: $\tau = 0$, which can be accomplished by a variance component score test. The mathematical derivation of the variance component score test can be found in **Supplementary Section 1.1** of the online **Supplementary Material**. In brief, the test statistic for $\boldsymbol{h} = \boldsymbol{0}$ in (**Eq. 8**) is

$$Q_1 = \left(\boldsymbol{y}^\star - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{W}\boldsymbol{K}^\star\boldsymbol{W}\left(\boldsymbol{y}^\star - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right), \tag{9}$$

where $\boldsymbol{y}^\star$ is a working response vector, $\boldsymbol{W}$ is a working weight matrix, and $\hat{\boldsymbol{\beta}}$ is the estimated coefficients under the null. For GLM, $\boldsymbol{y}^\star = \boldsymbol{D}_\pi(\boldsymbol{y} - \hat{\boldsymbol{\pi}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\boldsymbol{D}_\pi = \partial\boldsymbol{\eta}/\partial\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}}$ is a vector of fitted values returned by the null model $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. $\boldsymbol{W} = (\boldsymbol{D}_\pi\boldsymbol{V}_\pi\boldsymbol{D}_\pi)^{-1}$ and $\boldsymbol{V}_\pi$ is the variance-covariance matrix of the multinomial distribution evaluated at $\boldsymbol{\pi}$. For POM, $\boldsymbol{y}^\star = \boldsymbol{D}_\nu(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\nu}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\boldsymbol{D}_\nu = \partial\boldsymbol{\eta}/\partial\boldsymbol{\nu}$. $\boldsymbol{W} = (\boldsymbol{D}_\nu\boldsymbol{V}_\nu\boldsymbol{D}_\nu)^{-1}$, where $\boldsymbol{V}_\nu$ is the variance-covariance matrix of the cumulative probability $\boldsymbol{\nu}$.

For clustered study design, we write model (**Eq. 6**) in matrix notations

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{b} + \boldsymbol{h}, \tag{10}$$

where each component has three levels - category, cluster, and observation, except for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. Please refer to **Supplementary Section 1.2** of the online **Supplementary Material** for details of the model structure. Similarly, by applying pseudo-likelihood approach (Wolfinger and O'connell, 1993), the test statistic is

$$Q_2 = \left(\boldsymbol{y}^\star - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{\Sigma}^{-1}\boldsymbol{K}^\star\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}^\star - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right), \tag{11}$$

For GLMM, $\boldsymbol{y}^\star = \boldsymbol{D}_\pi(\boldsymbol{y} - \hat{\boldsymbol{\pi}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{U}\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{\pi}}$ is a vector of fitted values returned by the null model $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{b}$, and $\hat{\boldsymbol{\beta}}$ is a vector of estimated coefficients of the fix effect, $\hat{\boldsymbol{b}}$ is a vector of predicted values of $\boldsymbol{b}$. $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1} + \boldsymbol{U}\boldsymbol{G}^\star\boldsymbol{U}'$, where $\boldsymbol{W}^{-1} = \boldsymbol{D}_\pi\boldsymbol{V}_\pi\boldsymbol{D}_\pi$, and $\boldsymbol{G}^\star$ is a $(J-1)\times(J-1)$ block matrix with entries $\boldsymbol{I}_n \otimes \boldsymbol{G}_{jh}$, $j, h = 1, \ldots, J-1$. For POMM, $\boldsymbol{y}^\star = \boldsymbol{D}_\nu(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\nu}}) + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{U}\hat{\boldsymbol{b}}$, $\boldsymbol{W}^{-1} = \boldsymbol{D}_\nu\boldsymbol{V}_\nu\boldsymbol{D}_\nu$ and $\boldsymbol{G}^\star$ is a $(J-1)$ block diagonal matrix with entries $\boldsymbol{I}_n \otimes \boldsymbol{G}_{jj}$.

2.4 *p*-Value Calculation

While deriving the test statistics for Q_1 and Q_2 is relatively straightforward in the pseudo-likelihood framework (as detailed in **Supplementary Section 1** of the online **Supplementary Material**), obtaining their null distributions to calculate *p*-values is never an easy task. A major challenge lies in that classic asymptotic results in the likelihood framework tend to be inaccurate due to the relatively small sample size in microbiome studies (e.g., less than few hundred) and the over-dispersion in microbiome data (Chen et al., 2016). Small-sample correction procedures are available within relatively easier models such as the linear regression models or linear mixed model in literature (Chen et al., 2016; Zhan et al., 2017a; Zhan et al., 2018; Zhan et al., 2021). Yet, such an attempt in the more-complicated models (e.g., GLM, POM, GLMM, and POMM) considered in the current paper does not work out due to mathematical complexities of these models (e.g., canonical links are often unavailable or very complicated in such models). To this end, we resort to a pseudo-permutation strategy (Zhan et al., 2017b) to obtain accurate *p*-values in finite samples.

Briefly, the null distribution of all permutations of the test statistic can be approximated by the Pearson type III density, which is achieved by matching the first three moments. This strategy leads to a fast *p*-value calculation since we only need to use the matched Pearson type III density for *p*-value calculation without the need to draw real permutations (Zhan et al., 2017b). Essentially, we observe that the test statistics Q_1 and Q_2 can be reformulated as the trace of the product of two kernels matrix: a kernel matrix for outcomes ($\boldsymbol{K}_Y$) and a kernel matrix for microbiome data ($\boldsymbol{K}$ in **Eq. 7**). Here we still assume that the kernel matrix for microbiome data is identical across multiple categories. Therefore, we use $\boldsymbol{K}$ instead of the original $\boldsymbol{K}^\star = \boldsymbol{I}_{J-1} \otimes \boldsymbol{K}$ in test statistics Q_1 (**Eq. 9**) and Q_2 (**Eq. 11**). In the proposed framework, let the weighted residual $\boldsymbol{\epsilon} = \boldsymbol{W}(\boldsymbol{y}^\star - \boldsymbol{X}\boldsymbol{\beta})$ for independent data or $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^\star - \boldsymbol{X}\boldsymbol{\beta})$ for longitudinal data. The outcome kernel will be $\boldsymbol{K}_Y = \tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}'$, where $\tilde{\boldsymbol{\epsilon}} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_{J-1})$ is an $N \times (J-1)$ matrix, where $\boldsymbol{\epsilon}_j$ is the weighted residuals for the j-th category. Originally, $\boldsymbol{\epsilon} = Vec(\tilde{\boldsymbol{\epsilon}})$ is a vector of length $N(J-1)$, where $Vec(\cdot)$ denotes the operator that transforms a matrix into a column vector by vertically stacking the columns of the matrix. We refer the readers to previous publications for further details of *p*-values using the Pearson type III distribution (Zhan et al., 2017b).

Finally, recall that *p*-values of tests using different microbiome kernels could vary greatly depending on whether the kernel of choice captures the true underlying association pattern. To this end, we propose an omnibus test that first conducts individual tests using one of the kernels (Bray-Curtis, UniFrac, weighted UniFrac etc). And then combines these individual *p*-values (corresponding to different microbiome kernels) using the harmonic mean *p*-value (HMP) procedure (Wilson, 2019) for an omnibus *p*-value, based on which to conclude our inference of statistical association. This approach tends to be robust: it loses little power compared to when the best kernel (which is unknown in practice) is used and gains substantial power compared to when a poor choice of kernel is used.

3 RESULTS

3.1 Simulation Studies

3.1.1 Design of Simulations

We conducted comprehensive simulations to evaluate empirical type I error of MiRKAT-MC when there is no true associations, and statistical powers under different association patterns. For both independent and clustered study designs, microbiome compositions were simulated similarly as in previous studies (Zhao et al., 2015). Briefly, we first fitted a Dirchlet-multinomial distribution to a real upper-respiratory-tract microbiome dataset (Charlson et al., 2010), which contains 856 OTUs for 60 samples, and estimated the mean and dispersion parameters. We then used these estimated parameters to generate microbiome read counts via the Dirchlet-multinomial distribution. We intended to investigate what the most powerful kernel is when the causal OTUs are with or without phylogenetic relationships, and whether the abundance matters.

3.1.1.1 Independent Data

We considered simulations when there are three categories ($J = 3$) and when there are five categories ($J = 5$). Data from each sample was simulated independently, according to following model

$$\eta_{ji} = \alpha_j + 0.5 \times x_{i1} + 0.5 \times x_{i2} + \beta \times scale\left(\sum_{a \in \mathcal{A}} \boldsymbol{z}_{ia}\right), \quad (12)$$

where $i = 1, \ldots, N$ and $j = 1, \ldots, J-1$. We set the sample size $N =$ 80 or 200 for when $J = 3$, and $N = 150$ or 300 when $J = 5$. We simulated both nominal and ordinal outcomes, using appropriate link functions of η. For nominal data (GLM), $\alpha_j = -2$, and for ordinal data (POM), $\alpha_j = j - 4$. x_{i1} is a Bernoulli variable with probability of 0.5, whereas x_{i2} is a standard normal variable with mean 0 and variance 1. $\mathcal{A}$ is a set of outcome-associated OTUs among the p OTUs in the community. $\beta = 0$ for type I error simulations, for which the choice of $\mathcal{A}$ doesn't matter. *scale* is the operation that standardize the data to be mean 0 and variance 1 across all the samples.

For statistical power evaluation, we considered three scenarios. Under the first two scenarios, causal OTUs (in $\mathcal{A}$) were selected from clusters of related taxa on a phylogenetic tree. In specific, we first partitioned the simulated OTUs into 20 clusters through the partitioning-around-medoids (PAM) algorithm based on the corresponding phylogenetic tree. For scenario 1, we randomly chose a common cluster of the OTUs as the causal OTUs. For scenario 2, we chose the rarest cluster as the causal OTUs. For scenario 3, we picked the 10 most abundant OTUs without consideration of phylogenetic information. These three scenarios correspond to situations in which the weighted UniFrac, unweighted UniFrac and the Bray-Curtis distances are expected to be the most powerful, respectively. For scenarios 1 and 3, $\beta = 0.6, 0.8, 1.2, 1.6, 2.0$, and $\beta = 2, 4, 6, 8, 10$ for scenario 2.

For each scenario, we employed the weighted UniFrac (K_w), the unweighted UniFrac (K_u), the Bray-Curtis (K_{BC}) and a generalized UniFrac kernel with the parameter of 0.5 (K_5) for association testing. We also conducted the omnibus test by combining the p-values from all individual tests. To obtain convincing results, we generated 10,000 replicates to estimate the empirical type I errors and 2,000 replicates for statistical powers. Statistical significance was established under the nominal level of $\alpha = 0.05$ for all the simulation studies.

3.1.1.2 Clustered Data

We simulated two scenarios to assess MiRKAT-MC when data is clustered. We simulated a family based study and a longitudinal study. For family-based data, we included only a random intercept in the model to capture the correlation between samples, while for longitudinal data, both a random intercept and a random slope of time were involved in the model. We set the number of clusters $n = 30$ or 60 for three categories ($J = 3$), and $n = 50$ or 100 for five categories ($J = 5$). We simulated data under an unbalanced design: i.e., clusters may have a different number of observations. To achieve this, $n/2$ of the clusters have three observations and the other $n/2$ of the clusters have four observations. In this way, the total numbers of observations are $N = 105$ ($n = 30$) and $N = 210$ ($n = 60$) when $J = 3$ and $N = 175$ ($n = 50$) and $N = 350$ ($n = 100$) when $J = 5$. Within each cluster, the outcome category may vary over observations; e.g., in longitudinal studies, a person may be of one disease category at one time point and of a different disease category at a different time point.

The following model was utilized to simulate the data

$$\eta_{jik} = \alpha_j + 0.5 \times x_{ik1} + 0.5 \times x_{ik2} + \boldsymbol{u}_{ik}' \boldsymbol{b}_{ji} + \beta \times scale\left(\sum_{a \in \mathcal{A}} \boldsymbol{z}_{ika}\right), \quad (13)$$

where $i = 1, \ldots, n$, $j = 1, \ldots, J-1$, and $k = 1, \ldots, m_i$. The definition of the parameters η, α_j, β x_{ik1}, x_{ik2}, $\mathcal{A}$ and *scale* function are identical to the counterparts in model (**Eq. 12**). The same three scenarios of choices of $\mathcal{A}$ were considered for power assessment. When the model included only a random intercept, $u_{ik} = 1$ and b_{ji} was generated from $\sim \mathcal{N}(0, g_{jj})$, where $g_{jj} = \frac{1}{4}, 1, 4$ being the variance, respectively. When considering both a random intercept and a random slope of time, $\boldsymbol{u}_{ik} = (1, t_{ik})'$ and b_{ji} was simulated from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_{jj})$, where $\boldsymbol{G}_{jj} = \begin{pmatrix} g_{jj11} & g_{jj12} \\ g_{jj21} & g_{jj22} \end{pmatrix}$. We set $g_{jj11} = g_{jj22} = \frac{1}{4}, 1, 4$, respectively, and $g_{jj12} = g_{gg21}$ were determined by $\frac{1}{2} g_{jj11}$. Thus, the correlation between the random intercept and the random slope was fixed at $\frac{1}{2}$. The generation of random effect $\boldsymbol{b}_{ji}$ was different for GLMM and POMM. Specifically, for a fixed cluster i, for GLMM, we generated a new random vector of $\boldsymbol{b}_{ji}$ for each category j from the above distribution. For the ease of data generation, we kept $\boldsymbol{G}_{jj}$ the same across categories and did not consider correlation of $\boldsymbol{b}_{ji}$ between categories for nominal data. However, as we discussed in model (**Eq. 3**), GLMM enjoys the freedom of different $\boldsymbol{G}_{jj}$ and correlated $\boldsymbol{b}_{ji}$ across different categories. In contrast, for POMM, we generated a new random vector of $\boldsymbol{b}_i$ only once for each cluster i and then plugged the same $\boldsymbol{b}_i$ in model (**Eq. 13**) for different categories.

TABLE 1 | Empirical type I error rates of MiRKAT-MC for independent data with three-categories.

	MiRKAT-MCN		MiRKAT-MCO	
	***N* = 80**	***N* = 200**	***N* = 80**	***N* = 200**
K_w	0.0463	0.0465	0.0440	0.0470
K_u	0.0436	0.0491	0.0487	0.0492
K_{BC}	0.0488	0.0468	0.0469	0.0449
K_5	0.0479	0.0518	0.0476	0.0466
HMP	0.0502	0.0475	0.0461	0.0455

N *denotes the sample size.* K_w*, the weighted UniFrac kernel;* K_u*, the unweighted UniFrac kernel;* K_{BC}*, the Bray-Curtis kernel;* K_5*, the generalized UniFrac kernel with parameter 0.5; HMP, the omnibus test using harmonic mean* p*-value test.*

TABLE 2 | Empirical type I errors of MiRKAT-MC for clustered data with a random intercept and a random slope model with three-category outcomes.

	***n* = 30 (*N* = 105)**			***n* = 60 (*N* = 210)**		
g	**0.25**	**1**	**4**	**0.25**	**1**	**4**
MiRKAT-MCN						
K_w	0.0498	0.0492	0.0467	0.0478	0.0496	0.0484
K_u	0.0521	0.0533	0.0486	0.0449	0.0508	0.0478
K_{BC}	0.0519	0.0542	0.0494	0.0522	0.0478	0.0497
K_5	0.0527	0.0516	0.0521	0.0521	0.0468	0.0505
HMP	0.0514	0.0533	0.0472	0.0465	0.0478	0.0488
MiRKAT-MCO						
K_w	0.0500	0.0473	0.0474	0.0449	0.0498	0.0457
K_u	0.0486	0.0506	0.0487	0.0483	0.0483	0.0538
K_{BC}	0.0535	0.0507	0.0487	0.0453	0.0493	0.0485
K_5	0.0519	0.0471	0.0489	0.0476	0.0501	0.0486
HMP	0.0495	0.0467	0.0481	0.0452	0.0483	0.0475

n *indicates the number of clusters while* N *is the number of total observations.* g *denotes the variance of random effects. The definition of* K_w*,* K_u*,* K_{BC}*,* K_5*, and HMP is the same as* ***Table 1***.

3.1.2 Simulation Results

Empirical type I error rates of MiRKAT-MCN (for nominal outcomes) and MiRKAT-MCO (for ordinal outcomes) for independent data are reported in **Table 1**. As seen in the table, the empirical type I errors (at $\alpha = 0.05$) of MiRKAT-MC are all very close to the expected level. Empirical type I error rates under different mixed models for clustered data are reported in **Supplementary Tables S1–S4** (**Supplementary Section 2.1**, online **Supplementary Material**), which also show well-controlled type I errors for both nominal and ordinal outcomes.

Figure 1 shows the statistical powers of MiRKAT-MC using independent data with three categories. The results with five categories using independent data are in **Supplementary Figure S1** (**Supplementary Section 2.2**, online **Supplementary Material**). We observe that the tests with weighted UniFrac, unweighted UniFrac, and Bray-Curtis kernels are most powerful for scenarios 1, 2, and 3, respectively, regardless of whether the outcome is nominal or ordinal. However, the tests with Bray-Curtis kernel produced very little power in scenario 2, and the tests with unweighted UniFrac showed little power in scenario 3: the statistical power are close to their expected type I error. This is due to the differences in the true association signals that each of the kernels is designed to capture. The weighted UniFrac kernel is most powerful to capture signals that are dominated by common taxa in a cluster on a phylogenetic tree, while the unweighted UniFrac kernel shows its strengths when rare OTUs in a phylogenetic cluster determine the association (Chen et al., 2012). In contrast, the Bray-Curtis kernel is more appropriate when the outcome is associated with a set of OTUs with high abundance without referring to a phylogenetic tree. The Omnibus test considering all four kernels is robust. For example, among the tests using single kernels, only Bray-Curtis kernel shows significant powers under scenario 3. Yet, the omnibus test is still able to detect the association.

Table 2 shows the empirical type I error for our proposed methods when the data are clustered. Again, type I errors are well controlled to their nominal level. The statistical powers for simulations when data is clustered are presented in **Supplementary Figures S2–S5** (**Supplementary Section 2.2**, online **Supplementary Material**). Under three categories, **Supplementary Figure S2** corresponds to models with random intercepts, while **Supplementary Figure S3** presents models with both random intercepts and random slopes. Similarly, **Supplementary Figure S4** corresponds to models with random intercepts with five categories; **Supplementary Figure S5** is about models with both random intercepts and random slopes with five categories. The conclusions are similar to those of independent data. In addition, we observe that given a simulation scenario, a choice of kernel and an effect size, when the variance of the random effect (elements in $\boldsymbol{G}_{jj}$ in **Eq. 13**) increases, the statistical power decreases. It is because with the increase of the random effects, the within-cluster correlation increases, leading to a lower effective sample size.

3.2 Real Data Analysis

3.2.1 Associations Between Antibiotic Exposure and Gut Microbiome in Non-Obese Diabetic Mice in a Longitudinal Study

In the original study (Livanos et al., 2016), 555 non-obese diabetic mice were randomly assigned to three groups with each group exposed to distinct patterns and doses of antibiotics. The mice that were born to the same female and that were of the same sex constituted a cluster and each cluster received the same treatment. The first group (51 clusters, 203 mice) received sub-therapeutic continuous (STAT) antibiotic exposure, the second group (42 clusters, 167 mice) received therapeutic-dose pulsed (PAT) antibiotic exposure, and the last group (47 clusters, 135 mice) was not exposed to antibiotics and served as the control group (Hu et al., 2020). Microbiome data from fecal, cecal or ileal samples were collected longitudinally for each cluster by sacrificing a mouse, at 3, 6, 10, and 13 weeks from the start of the experiment (week 0). The number of observations per cluster varied from 2 (i.e., at week 3 and 6) to 4 (i.e., at week 3, 6, 10, and 13).

The goal of this application is to test the association between treatment groups (STAT, PAT or control) and gut microbiome. Here, we exclusively analyzed the fecal samples, leaving 499 samples from 140 clusters over time. The gut microbiome was

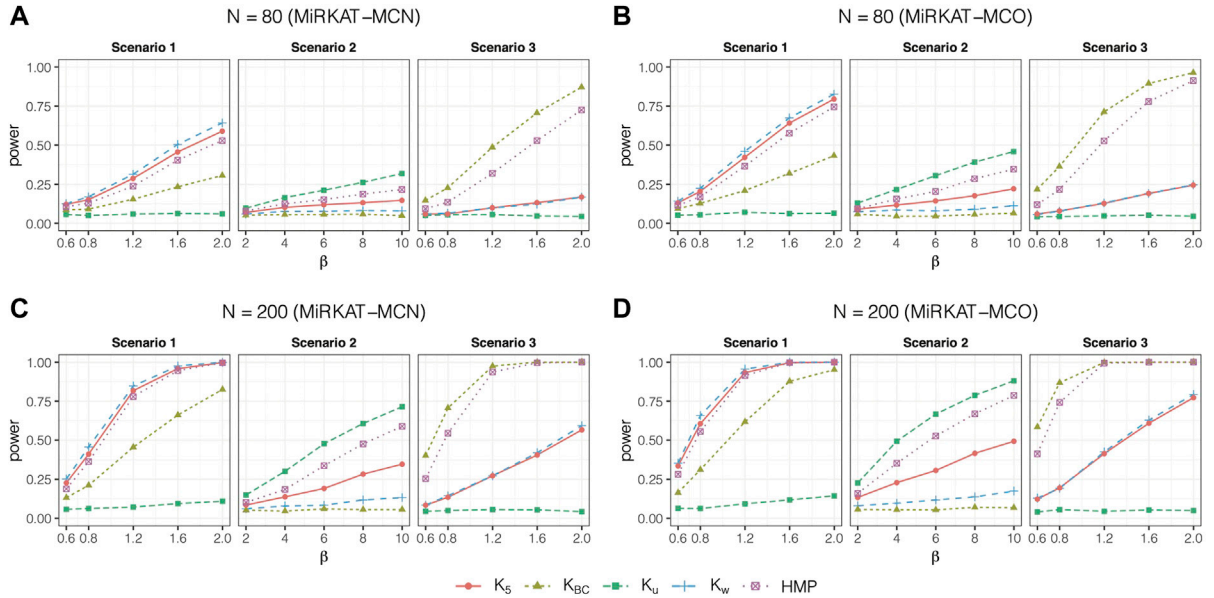


FIGURE 1 | Statistical powers of MiRKAT-MC for independent data with three categories. Scenario 1: $\mathcal{A}$ = A randomly selected common cluster among 20 clusters by PAM; Scenario 2: $\mathcal{A}$ = The rarest cluster among 20 clusters by PAM; Scenario 3: $\mathcal{A}$ = 10 most abundant OTUs. K_w, the weighted UniFrac kernel; K_u, the unweighted UniFrac kernel; K_{BC}, the Bray-Curtis kernel; K_5, the generalized UniFrac kernel with parameter 0.5; HMP, the omnibus test using harmonic mean *p*-value test. **(A)** MiRKAT-MCN with 80 total samples; **(B)** MiRKAT-MCO with 80 total samples; **(C)** MiRKAT-MCN with 200 total samples; **(D)** MiRKAT-MCO with 200 total samples.

profiled from each sample and the raw sequence data is available on the Qiita database (study ID 10508). Specifically, the V4 region of the bacterial 16S rRNA gene was PCR amplified, followed by performing paired-end sequencing of the amplicon library. We reprocessed the pre-joined and trimmed sequencing data through DADA2 pipeline in R (Callahan et al., 2016). As a result, the amplicon sequence variant (ASV) table was constructed. After removing chimeras identified by consensus across samples, the table contained 3031 ASVs. The ASV table was rarefied to an equal depth of 5,000 for each sample. We then assigned taxonomy based on Ribosomal Database Project's (RDP) training set 16, and constructed a phylogenetic tree using R package "phangorn" (Schliep, 2010). The tree was rooted by specifying the middle tip (i.e., 1515) as the outgroup. We calculated the UniFrac distance based on the rooted tree and the rarefied ASV table with the "GUniFrac" R package (Chen et al., 2012).

Here we first visually checked the relationship between gut microbiome composition and antibiotic treatment groups under different dissimilarity measures with PCoA plots (**Figure 2**). All 499 fecal samples are included in the plot, although they might be collected at different time points. Microbiome composition of the PAT group is clearly separated from that of the STAT group and that of the control group, under weighted UniFrac distance, generalized UniFrac distance and Bray-Curtis dissimilarity. However, under unweighted UniFrac distance, it is hard to distinguish the microbiome compositions of three treatment groups since they are clustered at two areas.

To show the performance of MiRKAT-MCN on independent nominal data, we selected samples at week 3 only. All 140 clusters had microbiome data available. By setting treatment groups as the dependent variable and adjusting for gender of mice, we observed very significant association between gut microbiome and the antibiotic treatment groups using weighted, unweighted, and generalized UniFrac kernels, Bray-Curtis kernel, and the omnibus test (all *p*-values < 0.0001). To better show the performance of the proposed model, and since the sample sizes of microbiome studies are usually smaller, we randomly subsampled 90 samples from the 140 samples at week 3. The down-sampled data consisted of 41 male and 49 female mice, and there were 36, 22, and 32 mice in the STAT, PAT and control groups, respectively. With the reduced sample size, all tests, including the tests using each of the kernels and the omnibus test, identified significant association between microbiome and antibiotic treament, with all *p*-values less than 0.0001, except for when using the unweighted kernel (*p*-value = 0.01).

We also applied MiRKAT-MCN for clustered data to this study. Similarly, we randomly selected 30 clusters with 105 samples (17 male and 13 female mice clusters) from the original dataset for analysis, where there were 15, 6, and 9 clusters in STAT, PAT, and control group, respectively. We applied MiRKAT-MCN for clustered data to evaluate the association between antibiotic treatment and microbiome, adjusting for sex and time (in weeks), and accounting for the cluster-specific correlation through a random intercept and a random slope of time. Again, we employed the same kernels as above and the omnibus test for analysis. Apart from the test using the unweighted UniFrac kernel with *p*-value only 0.03, all other tests were highly significant with *p*-values less than 0.001.

These two analyses indicate that antibiotic exposure during early life did alter the microbiome composition in non-obese

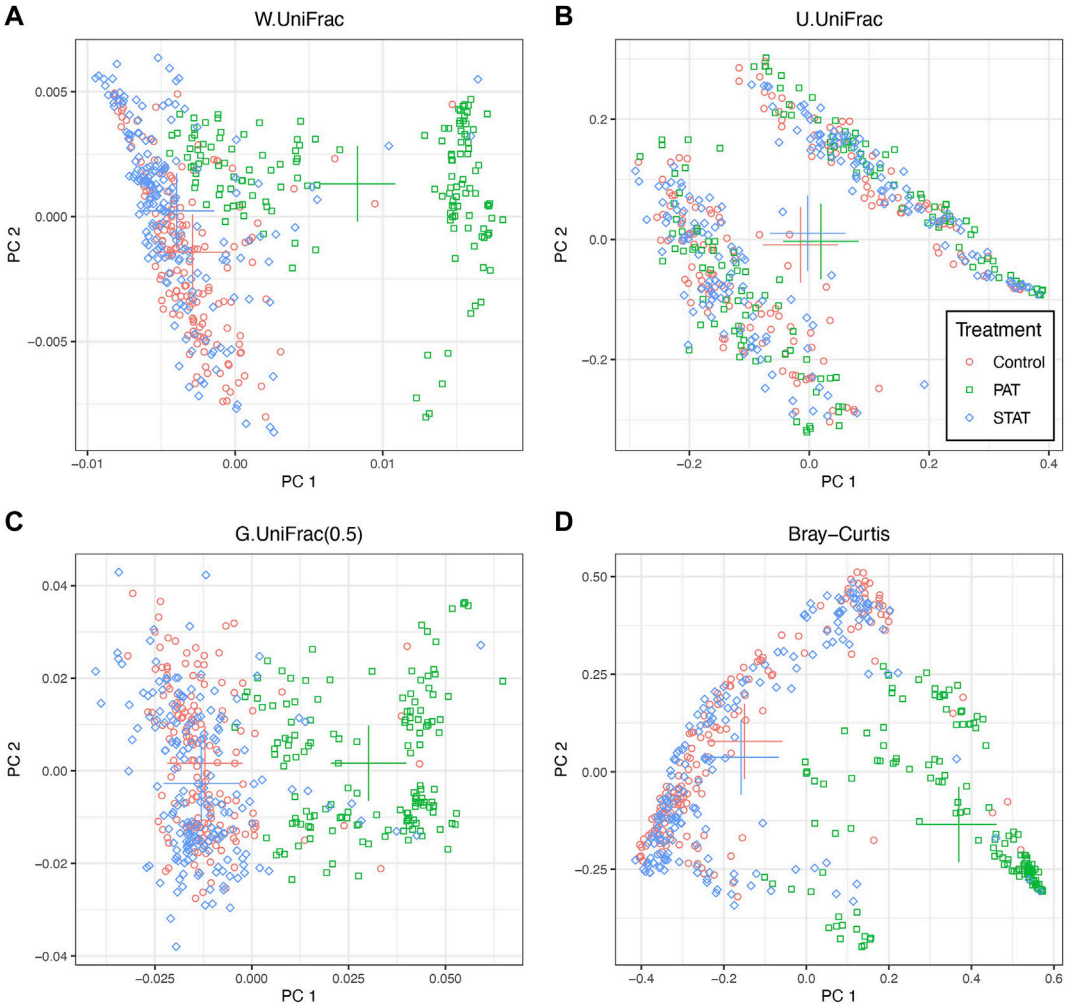


FIGURE 2 | The two-dimensional PCoA plots depicting microbiome composition for different antibiotic treatment groups under various dissimilarity measures. All 499 fecal samples are included in the plots. PAT, therapeutic-dose pulsed antibiotic exposure; STAT, sub-therapeutic continuous antibiotic exposure. The crosses denote the centroid of points of each treatment group. **(A)** W.UniFrac: weighted UniFrac distance; **(B)** U.UniFrac: unweighted UniFrac distance; **(C)** G.UniFrac(0.5): generalized UniFrac distance with tuning parameter a = 0.5; **(D)** Bray-Curtis: Bray-Curtis dissimilarity.

diabetic mice, no matter we stared at the week 3 or inspected over time. Moreover, the disparities of p-values by using different kernels, although all significant, suggest that the antibiotic use may have affected the relative abundance of OTUs, because the unweighted UniFrac kernel, which only accounts presence/absence of taxa and gives higher weight to rare taxa, provides the least significant result.

3.2.2 Associations Between Obesity and Gut Microbiome in a Family-Based Study

A study was conducted by Goodrich et al. (Goodrich et al., 2014) to investigate the role of host genetics on gut microbiome, and their impact on host phenotype, such as the body mass index (BMI). Fecal samples were collected from families in the United Kingdom. The V4 region of 16S rRNA gene was sequenced to identify the microbiome composition. The raw data was downloaded from the European Bioinformatics Institute (EBI) with accession numbers ERP006339 and ERP006342. We used QIIME (version 1.9.0-dev) (Caporaso et al., 2010) to assign the sequencing tags to 7,365 non-singleton OTUs at 97% similarity using the reference-based OTU-picking approach, and to generate a rooted phylogenetic tree. All samples were rarefied to 10,000 counts per sample before calculating the distance measures.

For this analysis, we focused on 311 samples from 145 monozygotic twin pairs. All the twins were female, aged from 27 to 83 with an median age of 63. In order to compare the performance of different methods, we treated the BMI as continuous, binary, three-category ordinal and three-category nominal data, and applied CSKAT (Zhan et al., 2018), GLMM-MiRKAT (Koh et al., 2019), MiRKAT-MCO and MiRKAT-MCN for each outcome type, respectively. CSKAT was developed for microbiome association analysis of clustered/longitudinal study for continuous outcomes while

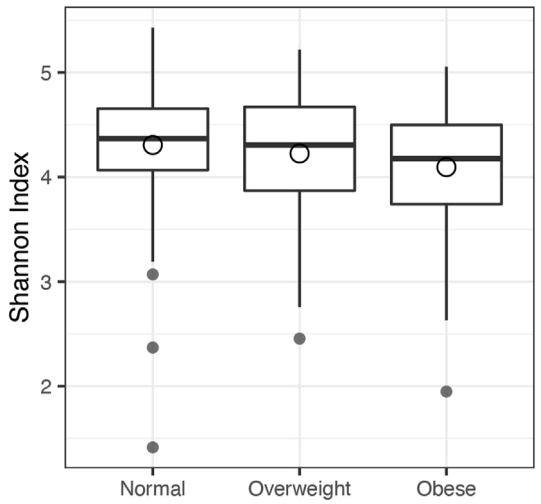


FIGURE 3 | The boxplot of Shannon index across BMI categories in United Kingdom twins study. Normal: BMI < 25; Overweight: 25 ≤ BMI < 30; Obese: BMI ≥30. The circle on each box denotes the mean of Shannon Index in that category.

TABLE 3 | *p*-values of testing for the BMI-microbiome association in United Kingdom twins dataset using different methods and kernels.

	CSKAT	GLMM-MiRKAT-Binary	MiRKAT-MCO	MiRKAT-MCN
K_w	0.1455	0.1750	0.2223	0.3268
K_u	0.0036	0.0182	**0.0014**	0.0033
K_{BC}	**0.0012**	0.0021	0.0016	0.0015
K_5	0.0278	0.0370	**0.0194**	0.0264
HMP	0.0036	0.0075	**0.0030**	0.0040

The bold value is the smallest significant p*-value across four methods given the kernel/method. The definition of K_w, K_u, K_{BC}, K_5, and HMP is the same as* ***Table 1****.*

GLMM-MiRKAT was for the similar association analysis for binary and count outcomes, respectively. For binary outcome, we classified the study participants into a non-obese (248 samples) and an obese group (63 samples) based on BMI < 30 or BMI ≥30. For the three-category outcome, we classify study participants into normal (BMI < 25), overweight (25 ≤ BMI < 30), and obese (BMI ≥30) groups, where there were 147, 101, and 63 samples in each group, respectively. We can treat the three categories as nominal or ordinal when applying MiRKAT-MC. For all the analyses, we assessed the microbiome-BMI (or BMI category) association, adjusting for age and including a twin-level random intercept to capture the within-twin-pair correlations due to common genetic, biological and other environmental factors. The weighted, unweighted, generalized UniFrac distance and the Bray-Curtis distance were used to construct kernel functions based on **Eq. 7**. The test statistics of CSKAT and GLMM-MiRKAT followed the original papers, but we used the same technique as MiRKAT-MC to calculate *p*-values, in order to ensure comparability.

TABLE 4 | Computation efficiency of MiRKAT-MC. Each result is the average time of one association test averaged from running 100 replicate association tests.

		MiRKAT-MCN (s)	MiRKAT-MCO (s)
Independent data			
$J = 3$	$N = 80$	0.0150	0.0139
	$N = 200$	0.0914	0.0796
$J = 5$	$N = 150$	0.0978	0.0426
	$N = 300$	0.7627	0.2568
Longitudinal data			
$J = 3$	$n = 30$ ($N = 105$)	6.438	2.844
	$n = 60$ ($N = 210$)	6.672	2.994
$J = 5$	$n = 50$ ($N = 175$)	11.964	4.758
	$n = 100$ ($N = 350$)	26.328	15.252

For longitudinal data, both random intercepts and random slopes of time are included in the null models. The weighted UniFrac kernel was applied without loss of generalization. n *denotes the number of clusters, whereas* N *is the total sample size. All the computation was conducted on a Macbook Pro (15-inch, 2019) laptop with 2.3 GHz 8-Core Intel Core i9 processor and 16 GB memory, without using parallel or other speed-up strategies.*

Figure 3 compares the microbiome Shannon index across the three BMI categories. The decreasing trend of Shannon index from the normal category to the obese category implies that higher BMI may reduce the microbiome diversity. The results of association analyses are shown in **Table 3**, where the smallest significant *p*-value of each kernel across four methods is bolded. At the first glance, all the individual tests provided significant association at type I error of 0.05 except when the weighted UniFrac kernel was used. The omnibus test also provided significant association. However, MiRKAT-MCO gave the smallest *p*-values when using the unweighted UniFrac, the generalized UniFrac and the omnibus test. MiRKAT-MCO was always more powerful than MiRKAT-MCN in this analysis, which is reasonable because MiRKAT-MCO utilized the order information in data. Both MiRKAT-MCO and MiRKAT-MCN were more powerful than GLMM-MiRKAT except when the weighted UniFrac kernel was used, for which none of the methods was significant. Our results are also consistent with the conclusion of the previous study (Zhan et al., 2018) that the unweighted UniFrac kernel and the Bray-Curtis kernel were most suitable for this dataset.

4 DISCUSSION

Multi-categorical outcomes, both nominal and ordinal, are increasingly common in biological and biomedical research over recent years. Investigating the subtle microbiome composition differences among multiple subtypes of a disease provides a broad view of microbiome variation. It is typically a first step to a further study of microbiome functionality and other related topics. Additionally, clustered designs, as a supplement to population-based studies, have become very popular recently when researchers are interested in dynamic variations or the variations among related individuals. While the toolbox for analyzing data collected from population-based studies is plentiful, methods for analyzing these clustered data are

underdeveloped. To fill these research gaps, we proposed MiRKAT-MC for testing for association between multi-categorical outcomes and microbial community compositions for both population-based and clustered/longitudinal studies.

Our major contributions in this paper are two-fold. First, we have successfully used the generalized logit model and the proportional odds model to enable direct association analysis between multi-categorical outcomes and microbiome compositions, without the need of combining categories or conducting pairwise comparisons. Existing approaches either compare two categories at a time and then conduct multiple testing correction, or combine multiple groups into a single category and compare it to the baseline. The pair-wise comparison approach tends to lose power due to the burden of multiple comparison. In addition, combining multiple groups into a single category can lead to substantial power loss when the microbiome effects on the categories are in opposite directions. However, when we have more than two categories, MiRKAT-MC can incorporate the heterogeneity in microbiome data and compare all non-reference categories to the reference category. Comparing to the potential alternative approach that first compares each pair of categories followed by multiple comparison adjustment, MiRKAT-MC would be much more powerful. Moreover, the new association analysis framework in the proportional odds model is extremely appealing for ordinal outcome data, as none of the existing approaches takes advantage of the order information in this particular type of data. Second, we have adapted a fast pseudo-permutation strategy previously developed under linear models to more complicated GLM(M) and POM(M) to achieve efficient and accurate p-values calculation. Unlike the ascendants which calculate p-values through either asymptotic distribution or direct permutation among exchangeable clusters, MiRKAT-MC controls type I error perfectly, even when the sample size is small, yet avoids the time-consuming and complex permutation.

As a non-parametric distance-based method, MiRKAT-MC comes with some limitations. First of all, the choice of distance metrics is subjective and could impact its performance. To this point, we propose to conduct analysis using multiple kernels/distances, generate multiple p-values and combine them via the harmonic mean approach (Wilson, 2019). Secondly, like other community level analysis of microbiome (Anderson, 2001; Zhao et al., 2015; Tang et al., 2016; Koh et al., 2019), MiRKAT-MC aggregates information across all taxa to form a community level test. This usually serves as the first step in understanding microbiome-phenotype relationship. However, these approaches do not provide insight on which taxa are driving the overall association. Thirdly, we used microbiome beta-diversity to define our distance/kernel matrix, which is convenient and proven useful. Many beta-diversities have been proposed and widely used in microbiome studies, which capture distinct characteristics of the underlying association pattern (see (Plantinga et al., 2017)). However, recent literature indicated that the structure of microbiome community may vary even when their diversities and compositions are comparable. In that context, if we are able to develop a sample-to-sample distance matrix that captures the important structure variations, such distance can be easily incorporated into our framework. Developing a kernel/distance for subtle structural differences in microbiome communities can be an interesting scientific endeavor, however, it is beyond the scope of this paper.

Computational efficiency of MiRKAT-MC is investigated and reported in **Table 4**. MiRKAT-MC is extremely fast when dealing with independent data. When data is clustered, the computational time increases substantially, mainly because of the increased time in fitting the null GLMM/POMM in the presence of random effects. Nevertheless, the computational time for MiRKAT-MC is very manageable even with clustered data. Given that most microbiome studies are relatively small in sample size, for three-category data, MiRKAT-MC can usually be accomplished in 0.1 s for population-based studies with sample size less than 200, and in 7 s for clustered studies with total sample size less than 210.

In summary, we propose MiRKAT-MC, a microbiome regression association test for multi-categorical outcomes with independent and clustered study designs. The proposed methods show well controlled type I errors and high power over multiple scenarios through extensive simulations and better performance than competitors in real data analyses. It is easy to use and fast to compute. We believe that MiRKAT-MC will enrich the toolbox of researchers to conduct microbiome research with multi-categorical outcomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZJ developed the method, conducted the simulation studies and real data applications, wrote the manuscript and the R program. MH and JC preprocessed the real data. XZ and NZ conceived the study and critically reviewed the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported in part by NIH for the Environmental Influences of Child Health Outcomes 531 (ECHO) Data Analysis Center (U24OD023382) and by Mayo Clinic Center for Individualized Medicine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.841764/full#supplementary-material

REFERENCES

Anderson, M. J. (2001). A New Method for Non-parametric Multivariate Analysis of Variance. *Austral Ecol.* 26, 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Bray, J. R., and Curtis, J. T. (1957). An Ordination of the upland forest Communities of Southern wisconsin. *Ecol. Monogr.* 27, 325–349. doi:10.2307/1942268

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, 581–583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). Qiime Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods* 7, 335–336. doi:10.1038/nmeth.f.303

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PloS one* 5, e15216. doi:10.1371/journal.pone.0015216

Chen, E. Z., and Li, H. (2016). A Two-Part Mixed-Effects Model for Analyzing Longitudinal Microbiome Compositional Data. *Bioinformatics* 32, 2611–2617. doi:10.1093/bioinformatics/btw308

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating Microbiome Composition with Environmental Covariates Using Generalized Unifrac Distances. *Bioinformatics* 28, 2106–2113. doi:10.1093/bioinformatics/bts342

Chen, J., Chen, W., Zhao, N., Wu, M. C., and Schaid, D. J. (2016). Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. *Genet. Epidemiol.* 40, 5–19. doi:10.1002/gepi.21934

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., et al. (2014). Temporal Variability Is a Personalized Feature of the Human Microbiome. *Genome Biol.* 15, 531. doi:10.1186/s13059-014-0531-y

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current Understanding of the Human Microbiome. *Nat. Med.* 24, 392–400. doi:10.1038/nm.4517

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., et al. (2014). Human Genetics Shape the Gut Microbiome. *Cell* 159, 789–799. doi:10.1016/j.cell.2014.09.053

He, Q., Liu, Y., Liu, M., Wu, M. C., and Hsu, L. (2021). Random Effect Based Tests for Multinomial Logistic Regression in Genetic Association Studies. *Genet. Epidemiol.* 45, 736–740. doi:10.1002/gepi.22427

Hu, J., Wang, C., Blaser, M. J., and Li, H. (2020). Joint Modeling of Zero-inflated Longitudinal Proportions and Time-to-event Data with Application to a Gut Microbiome Study. *Biometrics*. [Epub-ahead of prin]. doi:10.1111/biom.13515

Jiang, H., Ling, Z., Zhang, Y., Mao, H., Ma, Z., Yin, Y., et al. (2015). Altered Fecal Microbiota Composition in Patients with Major Depressive Disorder. *Brain Behav. Immun.* 48, 186–194. doi:10.1016/j.bbi.2015.03.016

Koh, H., Li, Y., Zhan, X., Chen, J., and Zhao, N. (2019). A Distance-Based Kernel Association Test Based on the Generalized Linear Mixed Model for Correlated Microbiome Studies. *Front. Genet.* 10, 458. doi:10.3389/fgene.2019.00458

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013a). Fusobacterium Nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host & Microbe* 14, 207–215. doi:10.1016/j.chom.2013.07.007

Kostic, A. D., Howitt, M. R., and Garrett, W. S. (2013b). Exploring Host-Microbiota Interactions in Animal Models and Humans. *Genes Dev.* 27, 701–718. doi:10.1101/gad.212522.112

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics* 63, 1079–1088. doi:10.1111/j.1541-0420.2007.00799.x

Liu, M., Liu, Y., Wu, M. C., Hsu, L., and He, Q. (2021). A Method for Subtype Analysis with Somatic Mutations. *Bioinformatics* 37, 50–56. doi:10.1093/bioinformatics/btaa1090

Livanos, A. E., Greiner, T. U., Vangay, P., Pathmasiri, W., Stewart, D., McRitchie, S., et al. (2016). Antibiotic-mediated Gut Microbiome Perturbation Accelerates Development of Type 1 Diabetes in Mice. *Nat. Microbiol.* 1, 16140. doi:10.1038/nmicrobiol.2016.140

Lozupone, C., and Knight, R. (2005). Unifrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi:10.1128/aem.71.12.8228-8235.2005

Naseribafrouei, A., Hestad, K., Avershina, E., Sekelja, M., Linløkken, A., Wilson, R., et al. (2014). Correlation between the Human Fecal Microbiota and Depression. *Neurogastroenterol. Motil.* 26, 1155–1162. doi:10.1111/nmo.12378

Ni, J., Shen, T. D., Chen, E. Z., Bittinger, K., Bailey, A., Roggiani, M., et al. (2017). A Role for Bacterial Urease in Gut Dysbiosis and Crohn's Disease. *Sci. Transl Med.* 9, eaah6888. doi:10.1126/scitranslmed.aah6888

Parikh, I. J., Estus, J. L., Zajac, D. J., Malik, M., Maldonado Weng, J., Tai, L. M., et al. (2020). Murine Gut Microbiome Association with Apoe Alleles. *Front. Immunol.* 11, 200. doi:10.3389/fimmu.2020.00200

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R., and Wu, M. C. (2017). MiRKAT-S: a Community-Level Test of Association between the Microbiota and Survival Times. *Microbiome* 5, 17. doi:10.1186/s40168-017-0239-9

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of Intestinal Prevotella Copri Correlates with Enhanced Susceptibility to Arthritis. *elife* 2, e01202. doi:10.7554/eLife.01202

Schirmer, M., Denson, L., Vlamakis, H., Franzosa, E. A., Thomas, S., Gotman, N. M., et al. (2018). Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course. *Cell Host & Microbe* 24, 600–610. e4. doi:10.1016/j.chom.2018.09.009

Schliep, K. P. (2010). Phangorn: Phylogenetic Analysis in R. *Bioinformatics* 27, 592–593. doi:10.1093/bioinformatics/btq706

Schloss, P. D. (2010). The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16s Rrna Gene-Based Studies. *Plos Comput. Biol.* 6, e1000844. doi:10.1371/journal.pcbi.1000844

Tang, Z.-Z., Chen, G., and Alekseyenko, A. V. (2016). Permanova-s: Association Test for Microbial Community Composition that Accommodates Confounders and Multiple Distances. *Bioinformatics* 32, 2618–2625. doi:10.1093/bioinformatics/btw311

Wilson, D. J. (2019). The Harmonic Mean P -value for Combining Dependent Tests. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1195–1200. doi:10.1073/pnas.1814092116

Wilson, N., Zhao, N., Zhan, X., Koh, H., Fu, W., Chen, J., et al. (2021). Mirkat: Kernel Machine Regression-Based Global Association Tests for the Microbiome. *Bioinformatics* 37, 1595–1597. doi:10.1093/bioinformatics/btaa951

Wolfinger, R., and O'connell, M. (1993). Generalized Linear Mixed Models a Pseudo-likelihood Approach. *J. Stat. Comput. Simulation* 48, 233–243. doi:10.1080/00949659308811554

Zhan, X., Banerjee, K., and Chen, J. (2021). Variant-set Association Test for Generalized Linear Mixed Model. *Genet. Epidemiol.* 45, 402–412. doi:10.1002/gepi.22378

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017a). A Small-Sample Multivariate Kernel Machine Test for Microbiome Association Studies. *Genet. Epidemiol.* 41, 210–220. doi:10.1002/gepi.22030

Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. (2017b). A Fast Small-sample Kernel independence Test for Microbiome Community-level Association Analysis. *Biom* 73, 1453–1463. doi:10.1111/biom.12684

Zhan, X., Xue, L., Zheng, H., Plantinga, A., Wu, M. C., Schaid, D. J., et al. (2018). A Small-sample Kernel Association Test for Correlated Data with Application to Microbiome Association Studies. *Genet. Epidemiol.* 42, 772–782. doi:10.1002/gepi.22160

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., et al. (2018). Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. *Front. Microbiol.* 9, 1683. doi:10.3389/fmicb.2018.01683

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in Microbiome-Profiling Studies with Mirkat, the Microbiome Regression-Based Kernel Association Test. *Am. J. Hum. Genet.* 96, 797–807. doi:10.1016/j.ajhg.2015.04.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 03 May 2022
doi: 10.3389/fgene.2022.847112

NISC: Neural Network-Imputation for Single-Cell RNA Sequencing and Cell Type Clustering

*Xiang Zhang[1,2], Zhuo Chen[1], Rahul Bhadani[1,3], Siyang Cao[3], Meng Lu[1], Nicholas Lytal[1,4], Yin Chen[5] and Lingling An[1,2,6]**

[1]Interdisciplinary Program in Statistics and Data Science, University of Arizona, Tucson, AZ, United States, [2]Department of Biosystems Engineering, University of Arizona, Tucson, AZ, United States, [3]Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, United States, [4]Department of Mathematics and Statistics, California State University at Chico, Chico, CA, United States, [5]College of Pharmacy, University of Arizona, Tucson, AZ, United States, [6]Department of Biostatistics and Epidemiology, University of Arizona, Tucson, AZ, United States

Single-cell RNA sequencing (scRNA-seq) reveals the transcriptome diversity in heterogeneous cell populations as it allows researchers to study gene expression at single-cell resolution. The latest advances in scRNA-seq technology have made it possible to profile tens of thousands of individual cells simultaneously. However, the technology also increases the number of missing values, i. e, dropouts, from technical constraints, such as amplification failure during the reverse transcription step. The resulting sparsity of scRNA-seq count data can be very high, with greater than 90% of data entries being zeros, which becomes an obstacle for clustering cell types. Current imputation methods are not robust in the case of high sparsity. In this study, we develop a Neural Network-based Imputation for scRNA-seq count data, NISC. It uses autoencoder, coupled with a weighted loss function and regularization, to correct the dropouts in scRNA-seq count data. A systematic evaluation shows that NISC is an effective imputation approach for handling sparse scRNA-seq count data, and its performance surpasses existing imputation methods in cell type identification.

Keywords: imputation, deep learning, single cell RNA-seq, dropout, autoencoder

OPEN ACCESS

Edited by:
Robert Friedman, Retired from University of South Carolina, United States

Reviewed by:
Andrea Tangherloni, University of Bergamo, Italy
Lu Zhang, Hong Kong Baptist University, Hong Kong SAR, China

****Correspondence:***
Lingling An
anling@email.arizona.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 January 2022*
Accepted: *04 April 2022*
Published: *03 May 2022*

Citation:
Zhang X, Chen Z, Bhadani R, Cao S, Lu M, Lytal N, Chen Y and An L (2022) NISC: Neural Network-Imputation for Single-Cell RNA Sequencing and Cell Type Clustering. Front. Genet. 13:847112. doi: 10.3389/fgene.2022.847112

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is designed to profile gene expression at the single-cell level, making it possible to study the heterogeneity among individual cells (Pierson and Yau, 2015). However, one important characteristic of scRNA-seq data is a phenomenon called "dropout", which causes challenges in data analysis. These dropout events occur because of the low amounts of genetic material in individual cells and inefficient mRNA capture, as well as the stochasticity of mRNA expression (Lin et al., 2017). Specifically, a large number of dropouts is due to transcripts lost in the RNA reverse transcription procedure during library preparation (Gordon et al., 2015). In other words, many zero counts in the gene expression data are not "true" values. Consequently, the scRNA-seq data may be incredibly sparse due to the high dropout rate, e.g., more than 90% of the expression counts have values of zero. Imputation has become an essential preprocessing step for downstream analysis of scRNA-seq data (Tracy et al., 2019). Recent studies have shown that some imputation methods improve downstream analysis and have already been implemented in scRNA-seq analysis pipelines (Zhang and Zhang, 2018). Meanwhile, with the increasing size of scRNA-seq data sets, appropriate imputation methods are necessary to compensate for these dropouts to reduce the impacts of missing values (Angerer et al., 2017).

Many methods have recently been developed for modeling and processing scRNA-seq count data, including scVI (Lopez et al., 2018), VASC (Wang and Gu, 2018), scSVA (Sun et al., 2019), scVAE (Gronbech et al., 2020), and scAEspy (Tangherloni et al., 2021), which used neural networks to reduce the noisy dimension to increase the accuracy of downstream analysis. There also exists quite a number of methods to impute the missing values in scRNA-seq data, including scImpute, MAGIC (Van Dijk et al., 2018), SAVER (Huang et al., 2018), DrImpute (Gong et al., 2018), VIPER (Chen and Zhou, 2018), ALRA (Linderman et al., 2018), EnImpute (Zhang et al., 2019) and scDoc (Ran et al., 2020). In ScImpute, separated Gamma-Normal mixture models are constructed for different cell subgroups to calculate the probabilities of drop-out. It leverages information of cell similarity in terms of genes with a lower dropout probability and then imputes the values of genes with higher dropout probability. MAGIC is a method that shares information across similar cells *via* data diffusion to predict the true gene expression level. SAVER is a Bayesian-based imputation method that imputes dropout values and generates a substitution for each gene. DrImpute is a clustering-based method that generates estimations using cluster priors and distance matrices. ALRA is an adaptively-thresholded low-rank approximation method that rescales the scRNA-seq expression matrix using randomized singular value decomposition. VIPER is a statistical method that fits a linear model for each cell by cell-cell interaction.

Basically, these methods impute dropouts by leveraging information on similarities between cells/genes using the correlation structure of the scRNA-seq data. For example, current imputation approaches, including scImpute and DrImpute, identify similar cells/genes based on clustering and then impute the missing data by averaging the gene expression values for each detected cluster. The accuracy of these imputation methods highly relies on clustering analysis. EnImpute combines the imputation results obtained from eight different imputation methods and calculates the expected values. scDoc imputes dropout events by leveraging information for the same gene from highly similar cells. However, current methods may fail to capture the nonlinearity and the count structure of the scRNA-seq data. Moreover, it becomes more challenging for the traditional imputation methods to handle datasets with increasing size (Eraslan et al., 2019).

Recently, some deep learning-based imputation methods have been developed for efficiently handling the higher dimensional scRNA-seq data, such as DCA (Eraslan et al., 2019), DeepImpute (Arisdakessian et al., 2019), AutoImpute (Talwar et al., 2018), LATE (Badsha et al., 2020), scIGAN (Xu et al., 2020), and scGNN (Wang et al., 2021). DCA is a neural network-based denoising method for scRNA-seq count data. This method assumes that the scRNA-seq count data follow a negative binomial distribution and then are denoised by maximizing a likelihood function. DeepImpute is a deep learning-based method that splits the genes into several subsets of neural networks. However, these imputation methods lack accuracy and power in handling highly sparse data. AutoImpute uses autoencoder with one hidden layer to impute missing values in scRNA-seq data by minimizing the Euclidean cost function. LATE uses autoencoder to train on nonzero data by minimizing the loss function, therefore imputing the missing values based on information of dependence between genes and cells. scIGAN uses generative adversarial networks for scRNA-seq imputation. scGNN uses a graph neural network for scRNA-seq analysis.

In this study, we develop a novel imputation method, Neural Network-based Imputation for scRNA-seq data (NISC) to improve cell type clustering. It is based on neural networks with a novel weighted loss function, coupled with regularizations. Through a series of simulation studies and real data analysis, NISC is compared with the other imputation methods, including AutoImpute, DCA, DeepImpute, LATE, SAVER, MAGIC, ScImpute, DrImpute, EnImpute, ALRA, VIPER, scDoc, scIGAN, and scGNN. The results show that NISC outperforms the existing imputation methods as it can recover the gene expression more correctly and distinguish the cell types more precisely, particularly for scRNA-seq data with high sparsity/noise.

2 METHODS

2.1 Neural Network Architecture

It is evident that the process of imputing the dropouts for scRNA-seq data is similar to the process of outlining a noisy image, so autoencoder is utilized to impute the sparse scRNA-seq data (Shao et al., 2013). Autoencoder is an unsupervised learning technique that has been used in image denoising (Vincent et al., 2010). The autoencoder technique allows nonlinear data vectors to be stacked, making the technique more powerful and able to learn complicated relations between layers (Mao et al., 2016). An autoencoder model consists of an encoder and a decoder. An encoder stage compresses the input data into a low-dimensional code, and then a similar decoder stage reconstructs the output data from the code (Hinton and Salakhutdinov, 2006). **Figure 1** shows the neural network architecture of NISC. The number of neurons for the hidden layer in the middle is usually much smaller than the number of neurons for the input/output layers to reduce the redundant information in data. In our method NISC, the number of neurons in the neural network architecture is set to be proportional to the number of genes.

2.2 Loss Function and Regularizations

It has been found that the main reason for dropouts in scRNA-seq data is due to failure of the reverse transcription of mRNA (Bengtsson et al., 2005; Reiter et al., 2011). Reverse transcription is an enzyme reaction; therefore, the Michaelis-Menten function can be used to model the relationship between dropout probability and gene expression for full-transcripts scRNA-seq data (Andrews and Hemberg, 2019). The following equation shows the dropout probability P_{ij} for the gene i in cell j using Michaelis-Menten kinetics (MMK) (Brennecke et al., 2013),

$$P_{ij} = 1 - \frac{S_{ij}}{K_M + S_{ij}} \tag{1}$$

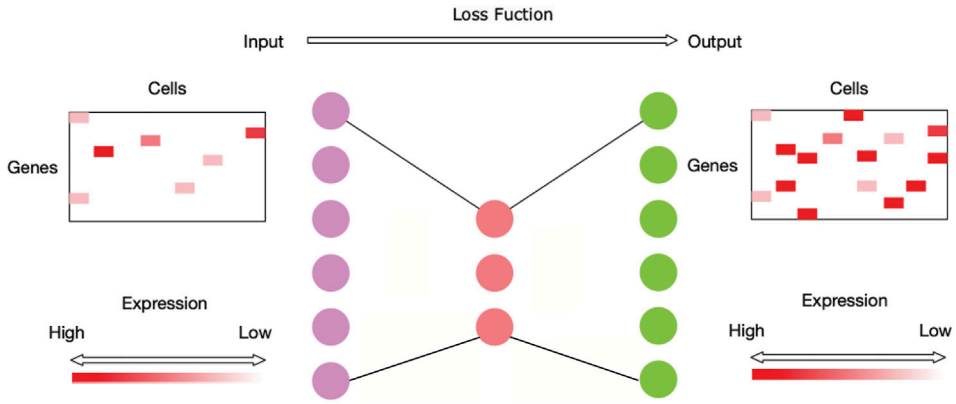


FIGURE 1 | Neural network architecture of NISC. This network is mainly composed of three hidden layers (i.e., dots with three different colors, purple, red, and green). The first hidden layer has neurons equal to twice the number of genes of the input data. It is followed by the second hidden layer in the middle with neurons equal to around half the number of genes of the input data. The third layer has neurons equal to the number of neurons of the first layer. This neural network is trained using an optimization process with a loss function to calculate the model error.

where S_{ij} is the observed gene expression level of gene i in cell j, and K_M is the Michaelis constant (Johnson and Goody, 2011). We use this probability to describe the dropout event, which will then be involved in the calculating the network's denoised output.

We propose a novel loss function with the mean square error weighted by the dropout probability estimated through Michaelis-Menten kinetics.

$$Loss = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(1 - P_{ij}\right)\cdot\left(\log\left(\hat{y}_{ij}\right) - \log\left(y_{ij}\right)\right)^2 + \alpha\cdot\left\|\beta\right\|_2 \qquad (2)$$

The loss function will be minimized through the autoencoder learning process. Note: the function "log" is the natural logarithm. The intuition behind this is that the estimated dropout probability P_{ij} affects the loss function adversely. In this manner, the imputed gene expression $\hat{y}_{ij}$ will be close to the observed gene expression y_{ij} when the estimated dropout probability P_{ij} is low. When we train an autoencoder network, a challenging problem is how to avoid overfitting. Overfitting refers to a neural network model that fits the training data too well to predict the pattern of new data. Overfitting is caused by noise in the training data, and the neural network includes this noise during the learning process. To avoid overfitting, we need to reduce the complexity of the network; therefore, we applied L_2 regularization (ridge regression) and dropout regularization to reduce the complexity of the autoencoder network (note: this is different from the term "dropout" event in scRNA-seq data). It is the first time that these two regularization techniques have been combined with an autoencoder network for imputation of scRNA-seq data. We define the regularization term $\|\beta\|_2$ as the L_2 norm of the weight matrix, that is, the sum of all squared weight values of the matrix (i.e., the first term in the above loss function). α is defined as the value of the regularization rate, which determines how powerful the effect of the regularization term will be. The regularization term $\|\beta\|_2$ is weighted by the scalar α and the regularization term will be excluded if α is zero. If α is too large, the neural network model will be less sensitive therefore increase the risk of underfitting. Conversely, if α is too small, the complexity of the model will be increased, so the risk of overfitting will be high. An appropriate value of α can be determined through cross-validation suggested by Ng et al.(2004).

In addition to L_2 regularization, dropout regularization is also used in NISC as it is a strategy to turn off neurons of the neural network with certain probability during training, which then further reduces the model's complexity (Srivastava et al., 2014). Furthermore, to mitigate the effect of reaching the local optimization peak by the neural network, the Adaptive Moment estimation algorithm is used to perform stochastic optimization (Eweda and Macchi, 1984).

2.3 Performance Evaluation

The proposed method is compared with the existing imputation methods through a series of simulated datasets and three real datasets. First, we visualize cell type sub-populations using 2-dimensional PCA (principal component analysis) plots or t-SNE (t-distributed stochastic neighbor embedding) plots (Kin et al., 2002; Kobak and Berens, 2019) depending on the data property (Anowar et al., 2021). UMAP (uniform manifold approximation) plots are also drawn (Becht et al., 2019). The commonly used unsupervised clustering algorithms, k-means (Na et al., 2010) and hierarchical clustering algorithms (Murtagh and Contreras, 2017), and Leiden algorithm(Traag et al., 2019), are used to group the cells on the reduced dimension of visualization results, which can then be used for calculating the performance measurements of each imputation method.

Four evaluation metrics are calculated to evaluate the accuracy of the cell type clusters in the visualization plots, including Adjusted Mutual Information (AMI) (Romano et al., 2014), Adjusted Rand Index (ARI) (Steinley, 2004), Fowlkes-Mallows Index (FMI) (Nemec and Brinkhurst, 1988), and Silhouette Score (SS) (Rousseeuw, 1987). Since we know the truth for the simulated data, the RMSE (Root Mean Square Error) is also

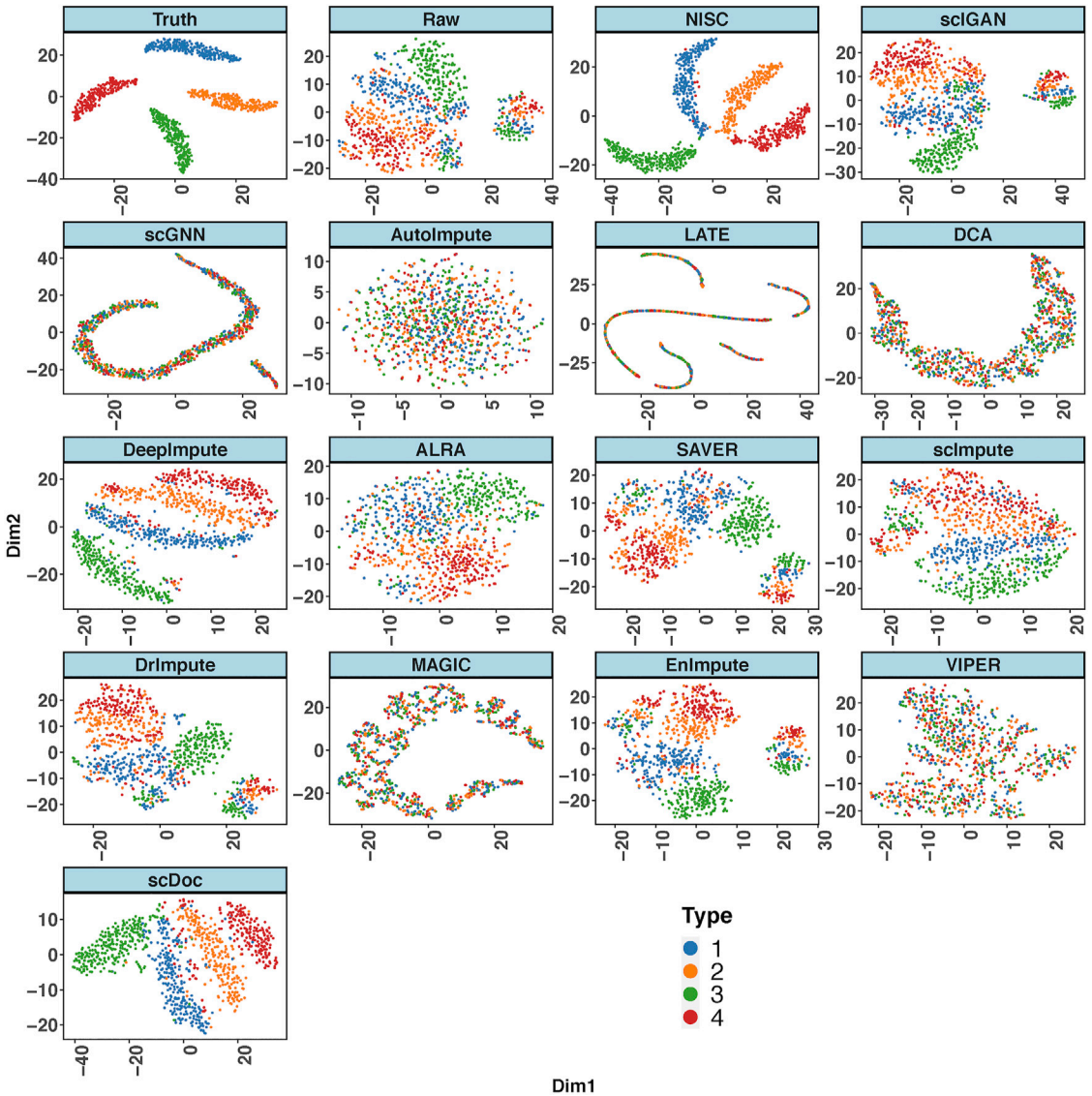


FIGURE 2 | NISC significantly improves the performance of t-SNE in visualizing simulated scRNA-seq count data. Plots of the first two components are calculated from the simulated ground truth data, raw data, and imputed data using various imputation methods. The dataset contains 800 genes and 1,000 cells in 4 cell types, with 90% sparsity. Cells are colored by cell types as indicated.

calculated between the imputed values and the truth to assess the performance of imputation methods (Blondel et al., 2008; Skinnider et al., 2019). Additionally, the heatmap of gene expression in the simulated studies is also drawn to demonstrate the direct comparison of the methods in detail.

3 RESULTS

3.1 NISC Enhances Cell Type Visualization in Simulated scRNA-Seq Data

To evaluate the performance of our imputation method, we compare it with existing methods on simulated scRNA-seq count data, which are generated by the widely used simulator, Splatter (Zappia et al., 2017). Both raw count data with dropouts/noise and its corresponding true data are available through simulations. The raw count data is the input data of the learning framework, and the ground truth data can be used to assess the performance of imputation. The count data are represented as an expression matrix, where each row is a gene, and each column is a cell. We consider three scenarios:

(1) Two cell types for 800 genes and 1,000 cells.
(2) Four cell types for 800 genes and 1,000 cells
(3) Four cell types for 2,000 genes and 10,000 cells

For each scenario, two sparsity levels are examined, i.e., approximately 80 vs 90%. In the Splatter simulation

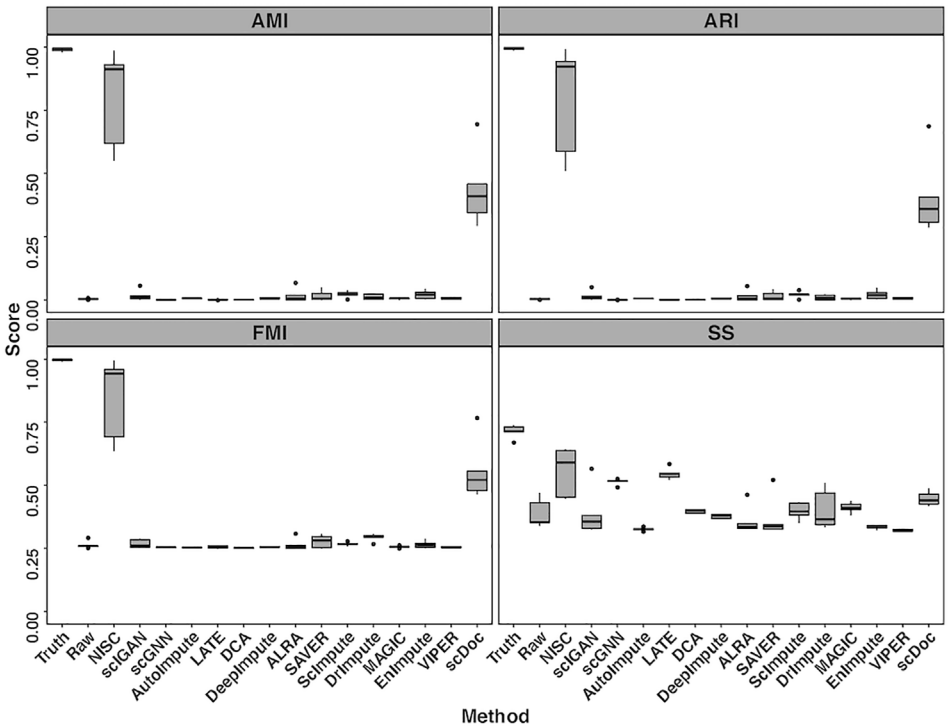


FIGURE 3 | Boxplots of four evaluation measures, including Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), Fowlkes-Mallows Index (FMI), and Silhouette Score (SS), are calculated for comparing NISC and other imputation methods. Each dataset contains 800 genes and 1,000 cells in 4 cell types, with 90% sparsity, and is replicated 10 times. Detailed information about these measurements can be found in the supplementary materials.

setting, the differential rate of 0.2 is used, indicating that 20% of the total genes are marker genes. As substantial noise is added to input data to mask cell type identities through simulation, our purpose is to predict the imputed values for the dropouts accurately and therefore identify cell types.

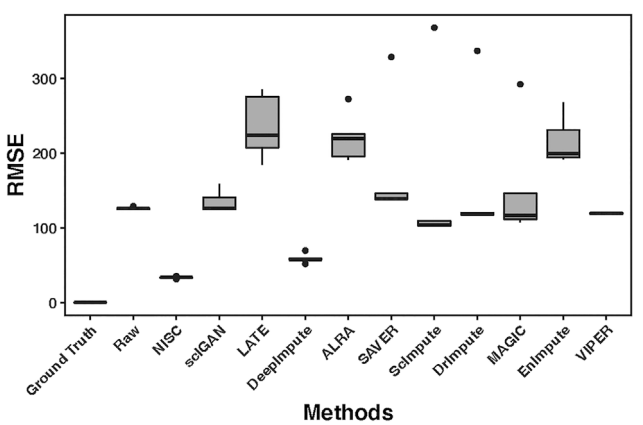


FIGURE 4 | RMSE (root mean square error) boxplots for the raw input and imputed data by each method. The RMSE is calculated between the ground truth and either the raw or imputed values. The raw dataset contains 800 genes and 1,000 cells in 4 cell types, with 90% sparsity, and is replicated 10 times. Detailed information about the RMSE can be found in the supplementary materials.

Our deep learning framework in NISC consists of three hidden layers with 1600, 400, and 1600 neurons, respectively, for the simulation data of 800 genes. For the case of 2,000 genes, the number of neurons for three hidden layers are 4,000, 1,000, and 4,000, respectively. A widely used active function, rectified linear unit (Xing et al., 2016), is employed to train each cell to capture the nonlinearity of the data. The number of neurons for the encoder/decoder layers is twice the number of genes, while the number of neurons for the hidden layer in the middle of the architecture is half of the number of genes. We compare NISC to other existing imputation methods in simulation data for various scenarios. The figures below are for the scenario (2). Some representative results for scenario 1) and 3) are included in the **Supplementary File**.

Figure 2 shows the t-SNE plots derived from the ground truth of cells, the raw input data, and the imputed data by NISC and other existing methods. The ground truth contains 4 cell types while the types are mixed in the raw data. This is due to the high sparsity (i.e., high noise, 90% data are zeros) in the raw input, which distorts the topology of the ground truth. NISC can accurately recover the dropouts, and the cells are clearly located in four groups/clusters, followed by scDoc and DeepImpute. However, it is challenging for other imputation approaches to distinguish the 4 cell types.

Four evaluation metrics, including AMI, ARI, FMI, and SS are calculated on the visualization result for the simulated data in **Figure 2**. To consider the data uncertainty (even with the same

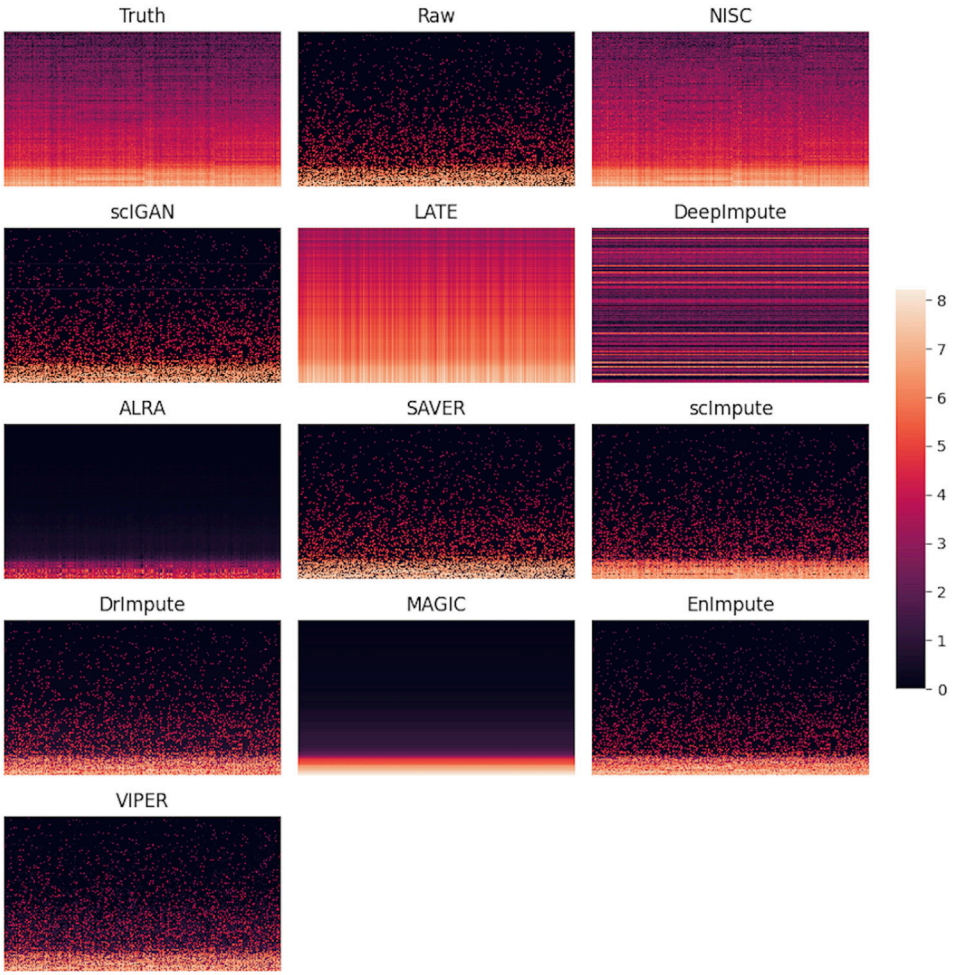


FIGURE 5 | Heatmaps of ground truth, raw simulated data, and imputed data by various methods. The simulated raw dataset contains 800 genes and 1,000 cells in 4 cell types, with 90% sparsity. Each row in the heatmap represents a gene, while each column represents a cell. The color bar shows the magnitude of the logarithm of gene expression values.

parameter settings) in the simulation, we generated ten replicates of datasets under each setting. **Figure 3** shows boxplots for four evaluation measures based on K-means clustering result of the t-SNE visualization. The boxplots of Leiden method are shown in **Supplementary Figure S2**. Higher values in measures indicate higher accuracy in cluster results. It is obvious that the performance of NISC surpasses all the existing imputation methods in clustering accuracy in this simulation study.

High accuracy in cell type visualization does not necessarily mean the imputed values are close to the true values. We calculated RMSE (root mean square error, the detailed definition can be found in the supplementary materials) between the ground truth value and the corresponding imputed value by each method. **Figure 4** shows boxplots of RMSE for 10 replicates of simulations. Compared with other imputation methods, the accuracy of NISC is highest, followed by DeepImpute, which is a neural network-based imputation method as well. Note: three imputation methods, DCA, AutoImpute and scGNN, are excluded from the RMSE plot as only highly variable genes are selected in these methods to perform imputation.

A direct comparison in gene expression values among the ground truth, raw data, and imputed data can be found in the heatmap plot (**Figure 5**). It shows that NISC imputed values are closest to the ground truth and therefore this method shows great capability in correcting the dropout values, which confirms the promising result in data visualization in **Figure 2**. Again, three imputation methods DCA, AutoImpute, and scGNN, are excluded from the heatmap plot as only highly variable genes are selected in these methods to perform imputation.

A consistent conclusion can be obtained from UMAP plot (**Supplementary Figure S3**) for this dataset. We also examine the impact of a different sparsity level (80%) on the imputation for the simulated data with 4 cell types and 2 cell types, respectively. When the sparsity of the simulated data with 4 cell types is about 80%, the cell populations can be revealed clearly in several imputation methods (**Supplementary Figure S4**), and NISC is one of them. Then, we observe that the performance of all

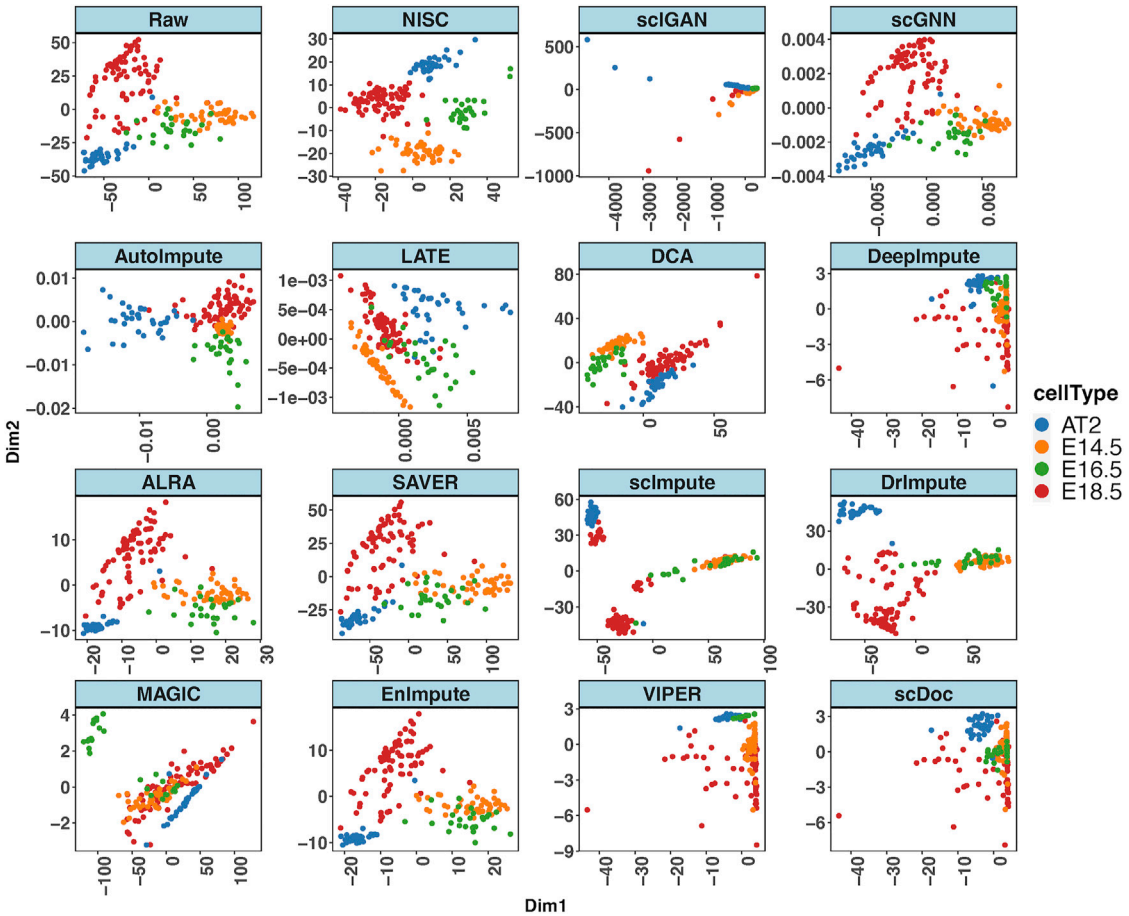


FIGURE 6 | NISC recovers the cell types (E14.5, E16.5, E18.5, and AT2) in mouse lung data. PCA plots of the raw data and imputed data by various imputation methods. The sparsity of the data is 72.6%. Cells are colored by cell types, which are reported in the original publication.

methods significantly decreases when dropout noise increases (**Supplementary Figure S4** vs **Figure 2**). A consistent conclusion can be obtained for the 2 cell types. **Supplementary Figure S5** shows an example of the t-SNE plot of 1,000 cells (in 2 cell types) and 800 genes with 80% sparsity. The cells are clearly separated into two groups/clusters by NISC, DeepImpute, DrImpute, EnImpute and scDoc, followed by scImpute, SAVER, and scIGAN.

For the case of 4 cell types with 10,000 cells, we only compared the deep-learning-based methods (**Supplementary Figure S6**). We noticed that the performances of three methods, NISC, DCA, and DeepImpute, are improved when the number of cells increases from 1,000 (**Figure 2**) to 10,000 (**Supplementary Figure S6**). The t-SNE plot in **Supplementary Figure S6** still shows that NISC surpasses other deep-learning-based methods, followed by DCA and DeepImpute.

Computational time: Among the deep-learning-based methods, LATE is the fastest, and scIGAN is the slowest. Specifically, the order of the computational time for seven deep learning-based methods is: LATE < DeepImpute < DCA < NISC < AutoImpute < scGNN < scIGAN. We used High Performance Computer systems with 2894 MHz CPU, 5 cores, and 36 GB memory on each core. For a simulation dataset with 2,000 genes and 10,000 cells, it took about 10 min for LATE, 12 h scIGAN, and 50 min for NISC.

3.2 NISC Improves Visualization Clarity and Clustering Accuracy in Real scRNA-Seq Data

3.2.1 Mouse Lung scRNA-Seq Data

We apply NISC and the compared methods on mouse lung scRNA-seq data (GSE52583) with 201 cells (Treutlein et al., 2014). **Figure 6** shows PCA plots for NISC and other imputation methods. The denoised data by imputation of scGNN, AutoImpute, ALRA, SAVER, scImpute, DrImpute, scDoc and EnImpute show E14.5 and E16.5 are not separated well, although cell type AT2 and E18.5 can be identified. In addition, with imputation of MAGIC, E16.5 is successfully identified, but E18.5, E14.5, and AT2 are mixed. By DCA, the 4 cell types (E14.5, E16.5, E18.5, and AT2) are grouped into two clusters, with two types in each. For DeepImpute, scIGAN and VIPER, the 4 cell types are mixed together. It seems that NISC can assign the 4 cell types into four clusters more accurately.

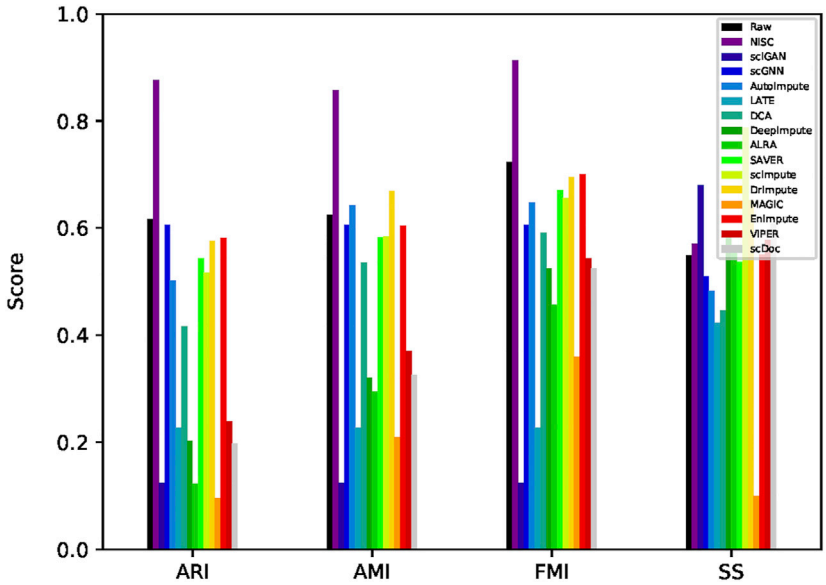


FIGURE 7 | Evaluation of clustering accuracy on the mouse lung data. Four measurements, AMI, ARI, FMI, and SS, are calculated for the imputed and raw data. The definitions of the measurements can be found in the supplementary materials.

The evaluation matrices on the clustering for this dataset are also calculated (**Figure 7**). Though NISC result does not provide the tightest clusters (from Silhouette score), among all the imputation methods, it scores the highest consistently across three measures of clustering accuracy, which confirms the separation pattern in the visualization in **Figure 6**.

3.2.2 Mouse Embryonic Data

We also apply NISC and the compared methods on scRNA-seq data of 92 mouse embryonic cells and 22,936 genes (GSE29087). The sparsity of the data is 83.04%. The cell types of this data set are reported in the original publication (Islam et al., 2011). We visualize the clustering result with t-SNE plots (**Supplementary Figure S7**), illustrating that, through NISC imputation, the 2 cell types, 48 mouse embryonic stem cells (ES) and 44 mouse embryonic fibroblasts (MEF), are separated, followed by DrImpute, DCA and scGNN. Through imputation of scIGAN, AutoImpute, LATE, ALRA, SAVER, MAGIC, EnImpute, SAVER, VIPER, and scDoc, the 2 cell types in this data are not separated well. With imputation of scGNN, DCA, and DrImpute, the 2 cell types are only somewhat separated. With scImpute, the cells are isolated into many tighter subclusters. In other words, some cells which should belong to the same cell type are scattered. The accuracy of clustering is assessed by four evaluation measures. Though NISC result does not provide the tightest clusters (from Silhouette score), among all the methods compared here, NISC is superior to others in terms of cluster accuracy ARI, AMI, and FMI. It improves the cluster results on original raw data.

3.2.3 Human Lung Adenocarcinoma Data

The above real scRNA-seq datasets do not have ground truth, since usually it is challenging to obtain the ground truth for real scRNA-seq data. Alternatively, it will be convincing to evaluate the performance of the imputation approaches if we use a real scRNA-seq dataset with low sparsity and distinct cell types and set it to be the ground truth data for evaluations. For this purpose, we apply the imputation methods on lung adenocarcinoma data (GSE69405) that profiles the gene expression of single cancer cells with TPM (normalization by transcripts per million) measurements (Soneson and Robinson, 2018). These cancer cells are originally from lung adenocarcinoma patient-derived xenograft (PDX) tumors, including four types, H358 human lung cancer cells (H358), cancer cells in PDX from primary tumors (LC-PT-45), an additional batch of PDX cells (LC-Pt-45-Re), and PDX cells for another lung cancer case (LC-MBT-15). This data set contains 176 cells, and the sparsity of the data is relatively low (46%). The cell types in this data can be clearly identified in the original data without imputation (**Figure 8A**). Therefore, we set the original data to be the ground truth. Following the method in (Arisdakessian et al., 2019) to generate noisy data, similarly, we mask the low-noise data by randomly changing some non-zeros to zeros so that the sparsity of the data is increased to 80% and the synthesized dataset here is termed as raw data.

T-SNE plots (**Figure 8A**) of the synthesized data show that NISC successfully recovers the cell types of the original data through imputing the sparse raw data. However, other imputation methods result in either one big cluster (i.e., all cells are mixed together) or several tight clusters, but each with two or more different cell types. A consistent conclusion can be obtained in evaluation plots (**Figure 8B**). Though the cells are not separately into tight clusters in NISC data, this method results in the highest cluster accuracy, considering the actual cell type status.

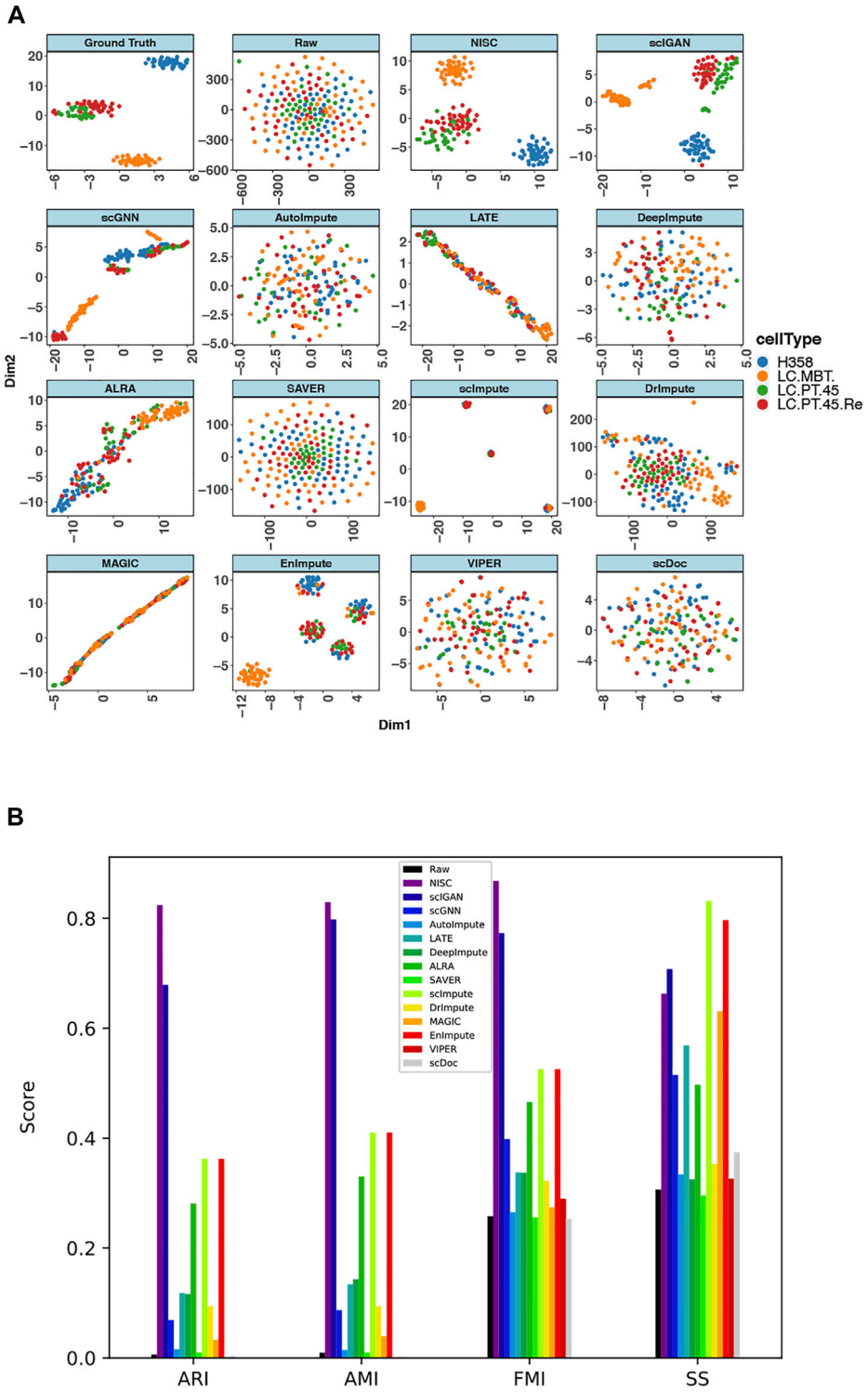


FIGURE 8 | NISC recovers the cell types in lung adenocarcinoma data (GSE69405) **(A)** plots of t-SNE components 1 and 2 derived from raw data, imputed data using NISC and other imputation methods. With additional zeros the sparsity of the data is 80%. Cells are colored by cell types, which are reported in the original publication **(B)** Bar plots of evaluation of cluster accuracy on the raw and imputed data. Four measurements, AMI, ARI, FMI, and SS, are calculated for the imputed and raw data. The definitions of the measurements can be found in the supplementary.

4 DISCUSSION

NISC is a data-driven method and does not require any prior knowledge. Real data and simulated data show that NISC can impute the dropouts in the scRNA-seq data, improving the accuracy of cell type clustering. Four performance measures were calculated to evaluate the clustering accuracy for the imputed data by various imputation methods. RMSE, which measures the distance between true (if available) and imputed values, was also calculated. Generally, compared with other existing estimation methods, NISC has a lower RMSE and a higher score in the evaluation measures of clustering accuracy.

NISC is an unsupervised neural network-based imputation method with autoencoder techniques implemented. Compared with other neural network-based methods, we investigated how different loss functions affect the imputation results. We developed a novel loss function weighted by Michaelis-Menten kinetics (MMK) and investigated its difference and standard mean square error (MSE) loss. Fig. S1 shows that the MMK loss can achieve more effective imputation under the sparse simulation setting, while by regular MSE the loss function is less effective. In addition, we add L2 regularization and dropout regularization to the model (Cortes et al., 2012) to avoid overfitting when denoising the input data. This is the first time the two regularizations are implemented simultaneously in the autoencoder model to impute scRNA-seq data.

An effective neural network for imputation requires sufficient neurons in the network. Due to many genes in scRNA-seq studies, GPUs are recommended for NISC to speed up the training process of the autoencoder network. NISC imputation is not suitable for some types of data which lose Michaelis-Menten kinetics, such as 10x Genomics data (Andrews and Hemberg, 2019), and some normalized data, for example, RPKM (Reads per kilo base per million mapped reads) or FPKM (Fragments Per Kilobase Million) (Lytal et al., 2020). However, TPM normalization is applicable as it maintains the data structure of the original gene expressions (Li and Li, 2018).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: github.com/anlingUA/NISC.

AUTHOR CONTRIBUTIONS

LA and XZ conceived the study. XZ and SC designed the methods and algorithms. XZ, ZC, RB, ML and NL performed the simulation studies. XZ, ZC, RB, YC and LA contributed to the real data analyses. XZ and LA drafted the manuscript. All authors revised, proofread, and approved the submitted manuscript.

FUNDING

This work has been partially supported by the National Institute of Health (1R01GM139829-01; 1P01AI148104-01A1; U19AG065169; 5P01AG052359-05) and the United States Department of Agriculture (ARZT-1361620-H22-149) to LA and by the National Institute of Health (R01AI149754 and R01ES027013) to YC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.847112/full#supplementary-material

REFERENCES

Andrews, T. S., and Hemberg, M. (2019). M3Drop: Dropout-Based Feature Selection for scRNASeq. *Bioinformatics* 35 (16), 2865–2867. doi:10.1093/bioinformatics/bty1044

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics. *Curr. Opin. Syst. Biol.* 4, 85–91. doi:10.1016/j.coisb.2017.07.004

Anowar, F., Sadaoui, S., and Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Comp. Sci. Rev.* 40, 100378. doi:10.1016/j.cosrev.2021.100378

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). DeepImpute: an Accurate, Fast, and Scalable Deep Neural Network Method to Impute Single-Cell RNA-Seq Data. *Genome Biol.* 20 (1), 211–214. doi:10.1186/s13059-019-1837-6

Badsha, M. B., Li, R., Liu, B., Li, Y. I., Xian, M., Banovich, N. E., et al. (2020). Imputation of Single-Cell Gene Expression with an Autoencoder Neural Network. *Quant Biol.* 8 (1), 78–94. doi:10.1007/s40484-019-0192-7

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* 37 (1), 38–44. doi:10.1038/nbt.4314

Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene Expression Profiling in Single Cells from the Pancreatic Islets of Langerhans Reveals Lognormal Distribution of mRNA Levels. *Genome Res.* 15 (10), 1388–1392. doi:10.1101/gr.3820805

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008 (10), P10008. doi:10.1088/1742-5468/2008/10/p10008

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for Technical Noise in Single-Cell RNA-Seq Experiments. *Nat. Methods* 10 (11), 1093–1095. doi:10.1038/nmeth.2645

Chen, M., and Zhou, X. (2018). VIPER: Variability-Preserving Imputation for Accurate Gene Expression Recovery in Single-Cell RNA Sequencing Studies. *Genome Biol.* 19 (1), 1–15. doi:10.1186/s13059-018-1575-1

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). *L2 Regularization for Learning Kernels*. arXiv preprint arXiv:1205.2653.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-Seq Denoising Using a Deep Count Autoencoder. *Nat. Commun.* 10 (1), 1–14. doi:10.1038/s41467-018-07931-2

Eweda, E., and Macchi, O. (1984). Convergence of an Adaptive Linear Estimation Algorithm. *IEEE Trans. Automat. Contr.* 29 (2), 119–127. doi:10.1109/tac.1984.1103463

Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: Imputing Dropout Events in Single Cell RNA Sequencing Data. *BMC bioinformatics* 19 (1), 1–10. doi:10.1186/s12859-018-2226-y

Gordon, A. J., Satory, D., Halliday, J. A., and Herman, C. (2015). Lost in Transcription: Transient Errors in Information Transfer. *Curr. Opin. Microbiol.* 24, 80–87. doi:10.1016/j.mib.2015.01.010

Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. (2020). scVAE: Variational Auto-Encoders for Single-Cell Gene Expression Data. *Bioinformatics* 36 (16), 4415–4422. doi:10.1093/bioinformatics/btaa293

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (5786), 504–507. doi:10.1126/science.1127647

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: Gene Expression Recovery for Single-Cell RNA Sequencing. *Nat. Methods* 15 (7), 539–542. doi:10.1038/s41592-018-0033-z

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., et al. (2011). Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq. *Genome Res.* 21 (7), 1160–1167. doi:10.1101/gr.110882.110

Johnson, K. A., and Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis-Menten Paper. *Biochemistry* 50 (39), 8264–8269. doi:10.1021/bi201284u

Kin, T., Tsuda, K., and Asai, K. (2002). Marginalized Kernels for RNA Sequence Data Analysis. *Genome Inform.* 13, 112–122.

Kobak, D., and Berens, P. (2019). The Art of Using T-SNE for Single-Cell Transcriptomics. *Nat. Commun.* 10 (1), 1–14. doi:10.1038/s41467-019-13056-x

Li, W. V., and Li, J. J. (2018). An Accurate and Robust Imputation Method scImpute for Single-Cell RNA-Seq Data. *Nat. Commun.* 9 (1), 997–999. doi:10.1038/s41467-018-03405-7

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and Accurate Clustering through Imputation for Single-Cell RNA-Seq Data. *Genome Biol.* 18 (1), 59–11. doi:10.1186/s13059-017-1188-0

Linderman, G. C., Zhao, J., and Kluger, Y. (2018). *Zero-preserving Imputation of scRNA-Seq Data Using Low-Rank Approximation*. BioRxiv, 397588.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep Generative Modeling for Single-Cell Transcriptomics. *Nat. Methods* 15 (12), 1053–1058. doi:10.1038/s41592-018-0229-2

Lytal, N., Ran, D., and An, L. (2020). Normalization Methods on Single-Cell RNA-Seq Data: an Empirical Survey. *Front. Genet.* 11, 41. doi:10.3389/fgene.2020.00041

Mao, X. J., Shen, C., and Yang, Y. B. (2016). *Image Restoration Using Convolutional Auto-Encoders with Symmetric Skip Connections*. arXiv preprint arXiv:1606.08921.

Murtagh, F., and Contreras, P. (2017). Algorithms for Hierarchical Clustering: an Overview, II. *Wiley Interdiscip. Rev. Data Mining Knowledge Discov.* 7 (6), e1219. doi:10.1002/widm.1219

Na, S., Xumin, L., and Yong, G. (2010). "Research on K-Means Clustering Algorithm: An Improved K-Means Clustering Algorithm," in Proceeding of the 2010 Third International Symposium on intelligent information technology and security informatics, Jian, China, April 2010 (IEEE), 63–67. doi:10.1109/iitsi.2010.74

Nemec, A. F. L., and Brinkhurst, R. O. (1988). The Fowlkes-Mallows Statistic and the Comparison of Two Independently Determined Dendrograms. *Can. J. Fish. Aquat. Sci.* 45 (6), 971–975. doi:10.1139/f88-119

Ng, A. Y. (2004). "Feature Selection, L 1 vs. L 2 Regularization, and Rotational Invariance," in Proceedings of the twenty-first international conference on Machine learning, July 2004, 78.

Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis. *Genome Biol.* 16 (1), 1–10. doi:10.1186/s13059-015-0805-z

Ran, D., Zhang, S., Lytal, N., and An, L. (2020). scDoc: Correcting Drop-Out Events in Single-Cell RNA-Seq Data. *Bioinformatics* 36 (15), 4233–4239. doi:10.1093/bioinformatics/btaa283

Reiter, M., Kirchner, B., Muller, H., Holzhauer, C., Mann, W., and Pfaffl, M. W. (2011). Quantification Noise in Single Cell Experiments. *Nucleic Acids Res.* 39 (18), e124. doi:10.1093/nar/gkr505

Romano, S., Bailey, J., Nguyen, V., and Verspoor, K. (2014). "Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance," in Proceedings of the International Conference on Machine Learning, Aug 2021, 1143–1151.

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Shao, L., Yan, R., Li, X., and Liu, Y. (2013). From Heuristic Optimization to Dictionary Learning: A Review and Comprehensive Comparison of Image Denoising Algorithms. *IEEE Trans. Cybern* 44 (7), 1001–1013. doi:10.1109/TCYB.2013.2278548

Skinnider, M. A., Squair, J. W., and Foster, L. J. (2019). Evaluating Measures of Association for Single-Cell Transcriptomics. *Nat. Methods* 16 (5), 381–386. doi:10.1038/s41592-019-0372-4

Soneson, C., and Robinson, M. D. (2018). Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis. *Nat. Methods* 15 (4), 255–261. doi:10.1038/nmeth.4612

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *J. machine Learn. Res.* 15 (1), 1929–1958. doi:10.5555/2627435.2670313

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychol. Methods* 9 (3), 386–396. doi:10.1037/1082-989x.9.3.386

Sun, S., Liu, Y., and Shang, X. (2019). "Deep Generative Autoencoder for Low-Dimensional Embeding Extraction from Single-Cell RNAseq Data," in Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, Nov. 2019 (IEEE), 1365–1372. doi:10.1109/bibm47256.2019.8983289

Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018). AutoImpute: Autoencoder Based Imputation of Single-Cell RNA-Seq Data. *Sci. Rep.* 8 (1), 1–11. doi:10.1038/s41598-018-34688-x

Tangherloni, A., Ricciuti, F., Besozzi, D., Liò, P., and Cvejic, A. (2021). Analysis of Single-Cell RNA Sequencing Data Based on Autoencoders. *BMC bioinformatics* 22 (1), 309–327. doi:10.1186/s12859-021-04150-3

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* 9 (1), 1–12. doi:10.1038/s41598-019-41695-z

Tracy, S., Yuan, G. C., and Dries, R. (2019). RESCUE: Imputing Dropout Events in Single-Cell RNA-Sequencing Data. *BMC bioinformatics* 20 (1), 1–11. doi:10.1186/s12859-019-2977-0

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing Lineage Hierarchies of the Distal Lung Epithelium Using Single-Cell RNA-Seq. *Nature* 509 (7500), 371–375. doi:10.1038/nature13173

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 174 (3), 716–729. doi:10.1016/j.cell.2018.05.061

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., and Bottou, L. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. machine Learn. Res.* 11 (12), 3371–3408.

Wang, D., and Gu, J. (2018). VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variational Autoencoder. *Genomics, proteomics & bioinformatics* 16 (5), 320–331. doi:10.1016/j.gpb.2018.08.003

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., et al. (2021). scGNN Is a Novel Graph Neural Network Framework for Single-Cell RNA-Seq Analyses. *Nat. Commun.* 12 (1), 1–11. doi:10.1038/s41467-021-22197-x

Xing, C., Ma, L., and Yang, X. (2016). Stacked Denoise Autoencoder Based Feature Extraction and Classification for Hyperspectral Images. *J. Sensors* 2016, 1–10. doi:10.1155/2016/3632943

Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z., and Zhou, X. (2020). scIGANs: Single-Cell RNA-Seq Imputation Using Generative Adversarial Networks. *Nucleic Acids Res.* 48 (15), e85. doi:10.1093/nar/gkaa506

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: Simulation of Single-Cell RNA Sequencing Data. *Genome Biol.* 18 (1), 1–15. doi:10.1186/s13059-017-1305-0

Zhang, L., and Zhang, S. (2018). Comparison of Computational Methods for Imputing Single-Cell RNA-Sequencing Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17 (2), 174–389. doi:10.1109/tcbb.2018.2848633

Zhang, X.-F., Ou-Yang, L., Yang, S., Zhao, X.-M., Hu, X., and Yan, H. (2019). EnImpute: Imputing Dropout Events in Single-Cell RNA-Sequencing Data via Ensemble Learning. *Bioinformatics* 35 (22), 4827–4829. doi:10.1093/bioinformatics/btz435

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

METHODS
published: 19 May 2022
doi: 10.3389/fgene.2022.846258

An Adaptive and Robust Test for Microbial Community Analysis

Qingyu Chen[1], *Shili Lin*[2]* *and Chi Song*[1]*

[1]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, United States, [2]Department of Statistics, College of Arts and Sciences, The Ohio State University, Columbus, OH, United States

In microbiome studies, researchers measure the abundance of each operational taxon unit (OTU) and are often interested in testing the association between the microbiota and the clinical outcome while conditional on certain covariates. Two types of approaches exists for this testing purpose: the OTU-level tests that assess the association between each OTU and the outcome, and the community-level tests that examine the microbial community all together. It is of considerable interest to develop methods that enjoy both the flexibility of OTU-level tests and the biological relevance of community-level tests. We proposed MiAF, a method that adaptively combines *p*-values from the OTU-level tests to construct a community-level test. By borrowing the flexibility of OTU-level tests, the proposed method has great potential to generate a series of community-level tests that suit a range of different microbiome profiles, while achieving the desirable high statistical power of community-level testing methods. Using simulation study and real data applications in a smoker throat microbiome study and a HIV patient stool microbiome study, we demonstrated that MiAF has comparable or better power than methods that are specifically designed for community-level tests. The proposed method also provides a natural heuristic taxa selection.

Keywords: human microbiome, association test, community-level test, OTU-level test, adaptive combination of p-values

OPEN ACCESS

Edited by:
Himel Mallick,
Merck, United States

Reviewed by:
Kalins Banerjee,
University of Michigan, United States
Siyuan Ma,
University of Pennsylvania, United States

***Correspondence:**
Shili Lin
shili@stat.osu.edu
Chi Song
song.1188@osu.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 December 2021*
Accepted: *28 March 2022*
Published: *19 May 2022*

Citation:
Chen Q, Lin S and Song C (2022) An Adaptive and Robust Test for Microbial Community Analysis. Front. Genet. 13:846258. doi: 10.3389/fgene.2022.846258

1 INTRODUCTION

Investigating the function of the microbiome in human health has become a burgeoning study field in recent years, which is attributed to the advent of new technologies for profiling complex microbial communities by 16 S rRNA gene sequencing (Lasken, 2012) or shotgun metagenomic sequencing (Hasan et al., 2014). Various microbial communities live throughout the human body and are associated with several diseases, such as colorectal cancer (Ahn et al., 2013), inflammatory bowel disease (Kostic et al., 2014) and obesity (Ley, 2010). Understanding the association between the microbiome and human disease may push back the frontiers of medical treatment.

Although the shotgun metagenomic sequencing enjoys higher resolution of taxonomic identification (Hasan et al., 2014), the reduced cost of 16 S rRNA gene sequencing makes it a more commonly used technology for microbiome studies to date. Using standard pipelines, 16 S sequences are clustered based on a prespecified similarity threshold (typically 97%) into operational taxonomic units (OTUs), each of which represents a taxonomic unit at a certain taxonomic rank, such as order, family, or genus (Nguyen et al., 2016). We note that some pipelines such as DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017) generate amplicon sequence variants (ASVs) instead of traditional OTUs. ASVs can be viewed as OTUs with the exact same sequences, and are sometimes referred as 100% OTUs. Because the analysis methods discussed here can be applied to

both OTUs and ASVs, we will not differentiate them and refer to both as OTUs in the rest of this paper. Since the initiation of Human Microbiome Project (Turnbaugh et al., 2007) in 2007, researchers have developed a variety of statistical methods to detect the possible association between microbiome diversity and an outcome of interest, such as a disease status.

There are two general categories of approaches for detecting associations—OTU-level methods and community-level methods. OTU-level methods test whether each individual OTU is associated with the outcome, while community-level methods test whether the microbial community in its entirety is associated with the outcome. Typically, the OTU-level methods test the association between a clinical outcome and the abundance of each OTU as a univariate covariate one-by-one. This univariate approach allows the development of many sophisticated OTU-level methods that can carefully accommodate the discrete and sparse nature of OTU-level abundance data. For example, QIIME (Caporaso et al., 2010), as a comprehensive pipeline, have the capability of performing OTU differential abundance tests using metagenomeSeq zero-inflated Gaussian (Paulson et al., 2013) and DESeq2 negative binomial Wald test (Love et al., 2014). The former developed a zero-inflated Gaussian distribution mixture model to avoid biases due to undersampling of the microbial community, while implementing a normalization method to deal with uneven sequencing depth. The later adapted the negative binomial model that has been popular in gene differential expression study to analyze microbiome data. In addition, QIIME also contains several classic statistical tests, such as ANOVA, Kruskal-Wallis, G-test, Mann-Whitney test, as well as the parametric and nonparametric t-test.

In practice, it is frequently more biologically relevant to perform community-level analysis, which jointly tests the association between a clinical outcome and a microbial community as a whole. These methods are often based on alpha diversity or beta diversity. Alpha diversity characterizes the complexity of the microbial community within each sample. Among them, the Inverse Simpson Diversity (Simpson, 1949), Shannon Indexes (Shannon, 1948) and Faith's phylogenetic diversity that incorporates phylogenetic relationships (Faith, 1992) are some of the most popular choices. After summarizing the complexity of the microbial community into a single alpha diversity metric, univariate methods such as regression models can be applied to detect the possible association between the alpha diversity and the clinical outcome. Adaptive microbiome α-diversity-based association analysis (aMiAD) (Koh, 2018) used the minimum p-value from association analyses based on different alpha diversity metrics as its test statistic, and assessed the p-value of the proposed test *via* a residual-based permutation method. Beta diversity, on the other hand, measures the distance or dissimilarity between each pair of biological samples. For example, Bray-Curtis dissimilarity measures the differences between two microbial communities by quantifying the non-overlapping OTU abundances (Bray and Curtis, 1957). Jaccard distance can be viewed as an "unweighted" version of Bray-Curtis dissimilarity, since it only relies on the presence or absence of OTUs without taking abundance information into account (Jaccard, 1901, Jaccard, 1912). Among many available distance metrics, the UniFrac distance incorporating phylogenetic information is one of the most popular metrics (Lozupone et al., 2007). It calculates the fraction of sums of branch lengths with their corresponding taxa only in one sample to both samples. Both weighted and unweighted versions of UniFrac are commonly used in microbial ecology, where the former accounts for abundance information of the taxa, while the latter only considers their presence or absence. Moreover, generalized UniFrac distances were proposed as a series of distance metrics—from unweighted to weighted UniFrac by assigning different weights on the branches (Chen et al., 2012). Based on the beta diversity or a distance metric, various community-level association testing methods have been proposed. Permutational Multivariate Analysis of Variance (PERMANOVA) (McArdle and Anderson, 2001), one of the pioneer community-level tests, is a non-parametric method that fits multivariate models for microbial community data to test whether the samples significantly differ across a categorical factor. It bears some resemblance to ANOVA but operates on a dissimilarity matrix and assesses p-values based on permutation. However, PERMANOVA usually adopts only one of the many available distance metrics with no confounder adjustment and cannot easily accommodate continuous traits (unless categorized arbitrarily). Microbiome Regression-based Kernel Association Test (MiRKAT) (Zhao et al., 2015), a more comprehensive method, was proposed to extend the outcome of interest to the continuous case. The phylogenetic dissimilarity matrix is transformed into a kernel matrix which measures the similarity of microbial communities between samples. MiRKAT regresses the clinical outcome on this semiparametric kernel machine while adjusting for potential confounders. It should be noted that MiRKAT is equivalent to PERMANOVA when no covariates are included. Besides, MiRKAT can combine multiple distance metrics by selecting the one that generates the smallest p-value.

Although OTU-level methods and community-level methods tackle the association testing problem from different angels, they are in fact related to each other. The statement that the microbial community is associated with the clinical outcome is equivalent to that at least one of the OTUs differs across the outcome status. Therefore, theoretically, the results of all the OTU-level tests can be summarized across the observed taxon units to draw a community-level conclusion about whether the microbial community is associated with the clinical outcome. Considering the vast availability of univariate models for different study designs that can be directly applied to OTU-level analysis, as well as the sophisticated OTU-level methods that accommodate unique aspects of microbiome data, it would be beneficial to combine them into community-level tests.

However, simply putting all OTU-level tests together without proper weighting or OTU selection will suffer from power loss, because not all OTUs may be associated with the outcome, and as thus, a naive combination may accumulate noises that eventually surpass association signals. Moreover, the number or proportion of OTUs that are not associated with the outcome is often unknown in practice. In contrast, adaptively and wisely assigning weights to the taxon units according to their

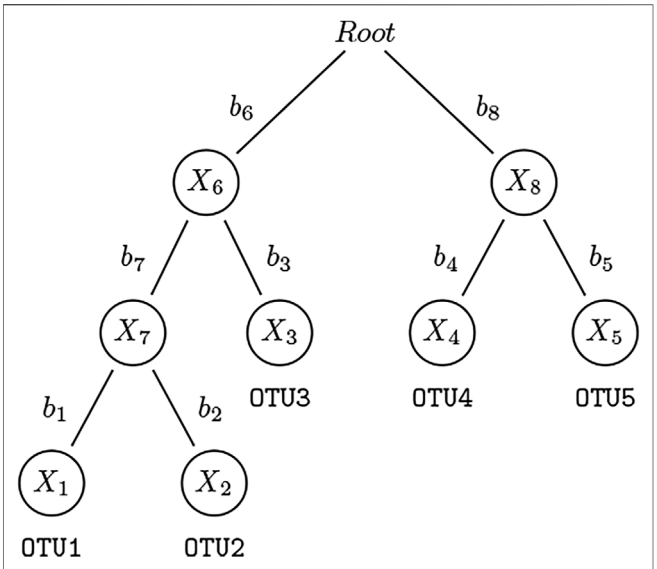


FIGURE 1 | An example of a rooted phylogenetic tree. This is a simple rooted phylogenetic tree containing 5 leaf nodes and 4 internal nodes. The counts for leaf nodes—OTU1–5 — can be obtained *via* standard pipelines. We assign the counts to internal nodes by summing up the counts of their children and refer to them also as "OTUs": For three of the four abundance representations used in this paper, OTU7 = OTU1 + OTU2, OTU6 = OTU7 + OTU3, and OTU8 = OTU4 + OTU5. We construct the extended OTU matrix containing all nodes (except the root) in this phylogenetic tree.

importance is a key to achieving greater statistical power. Some efforts have already been put into this area. For example, adaptive Microbiome-based Sum of Powered Score (aMiSPU) test (Wu et al., 2016) extended the aSPU test (Pan et al., 2014) to accommodate unique features of microbial data. This method adaptively combines the score statistics for two versions of generalized taxon proportions and resembles MiRKAT with weighted and unweighted UniFrac kernel. OMiAT (Koh et al., 2017) combines aSPU and MiRKAT by taking the minimum p-value from all the score tests of the two methods. aSPU used in OMiAT implements on standard compositional microbial data without incorporating phylogenetic information. However, the requirement of score statistic for taxon units in aMiSPU and aSPU may limit their applicability to different study designs where the score statistics may not be readily available. This requirement also makes aMiSPU and OMiAT inflexible to combine more sophisticated OTU-level testing methods that are specifically designed for microbiome data. Compared to score statistic, p-value is a more universally available statistic in OTU-level association tests, thus making it a more suitable target to combine, for the sake of flexibility. MiHC (Koh and Zhao, 2020), adapted from higher criticism test which aims to detect highly sparse signals, was tailored to accommodate different sparsity levels and incorporate phylogenetic information. It was more powerful for sparse microbial association signals than abundant ones. In this paper, inspired by Adaptive Fisher (AF) method (Song et al., 2016), we propose a p-value combination approach, Microbiome Adaptive Fisher method (MiAF), to aggregate p-values of OTU-level tests into a novel community-level association test. It should be noted that the focus of MiAF is to test whether the OTU community is associated with the outcome, instead of estimating the parameters of the association model. We compare the performance of MiAF to methods specifically designed for detecting community-level associations, and demonstrates comparable or better power for MiAF. We also discuss the potential of MiAF as a general p-value combination framework for microbial community-level tests under various study designs.

2 MATERIALS AND METHODS

2.1 Statistical Model and OTU-Level Tests

Suppose n subjects are observed and their microbial communities are profiled. For the ith subject, Y_i denotes the outcome of interest which can be binary or continuous, and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ic})$ denotes c covariates such as age and gender that are potentially associated with both the clinical outcome and microbial community, which we need to adjust for as potential confounders. We construct an "extended" OTU table containing all nodes (terminal and internal) in the phylogenetic tree. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im})$ be the counts of "extended" OTUs which consist of both leaf nodes and internal nodes (except for root node) for subject i, where m is the total number of "extended" OTUs. The count of an internal node is derived by summing up all the counts of the leaf node OTUs belonging to this taxon (see **Figure 1** for an illustration). Note that our method is not limited to bifurcating phylogenetic trees, it is applicable to multifurcating trees. The relative abundance of extended OTU k, $k = 1, \ldots, m$, in subject i, $i = 1, \ldots, n$, is $A_{ik} = X_{ik}/\sum_{j=1}^{q} X_{ij}$, where q is the number of leaf nodes, and the X_{ij}'s are arranged such that the first q entries in $\boldsymbol{X}_i$ are the leaf nodes in the same order for all individuals.

OTU abundance varies greatly in a microbial community. Some microbes are dominant, but most are rare. In practice, the underlying association patterns are unknown a priori. We do not have the knowledge of the characteristics of the truly associated OTUs nor their phylogenetic relationships that are captured by phylogenetic trees. Therefore, we incline to integrate the abundance information and phylogenic relationships adaptively to achieve a robust test under diverse underlying situations. When the associated OTUs are indeed phylogenetically related, incorporating phylogenetic information may boost the performance of an association analysis to a great extent. To accommodate such a situation, we define unweighted and weighted taxon proportions as $M_{ik}^{u} = I(A_{ik} > 0)$ and $M_{ik}^{w} = A_{ik}$ respectively for "extended" OTU k, $k = 1, \ldots, m$. The unweighted taxon proportion only considers the presence or absence of an OTU, whereas the weighted one takes the magnitude of the abundance information into account. Inspired by the generalized UniFrac distance metric (Chen et al., 2012), we also define a square-root transformed taxon proportion to attenuate the contribution by highly abundant OTUs as $M_{ik}^{.5} = A_{ik}{}^{0.5}$.

We also consider a taxon proportion restricted to leaf nodes only for situations where the associated OTUs are not phylogenetically related, since incorporating phylogenetic

information in this scenario may adversely affect testing performance. That is, we only include the weighted taxon proportion of leaf nodes in original OTU table defined as $M^a_{ik} = A_{ik}$, $k = 1, \ldots, q$.

We use the following generalized linear model to depict the association between the compositions of microbes in a community and the health outcome taking confounding covariates into consideration:

$$h(E[Y_i]) = \alpha_0 + \boldsymbol{Z}_i\boldsymbol{\alpha} + \sum_{k=1}^{d} M_{ik}\beta_k, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_c)^\top$ represents the effects of the c covariates, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_d)^\top$ are the effects of the OTUs, and M_{ik} can be any of the four abundance representations defined above (M^u_{ik}, M^w_{ik}, $M^{.5}_{ik}$ or M^a_{ik}); thus $d = q$ for M^a_{ik} and $d = m$ for the other three measures. Finally, $h(\cdot)$ is the link function, which is the logit function for binary outcomes or the identity function for continuous outcomes.

We are interested in determining whether there is an association between the outcome of interest and any OTU, which is equivalent to testing the following hypotheses:

$$H_0\colon \boldsymbol{\beta} = \boldsymbol{0} \text{ vs. } H_1\colon \boldsymbol{\beta} \neq \boldsymbol{0}.$$

The score statistics $\boldsymbol{U} = (U_1, \ldots, U_d)$ for $\boldsymbol{\beta}$ can be calculated as $\boldsymbol{U} = \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\boldsymbol{M}_i - \widehat{\boldsymbol{M}_i})$, where $\hat{\mu}_i$ is the expectation of Y_i under H_0, and $\widehat{\boldsymbol{M}_i} = (\widehat{M_{i1}}, \ldots \widehat{M_{id}})$ are the fitted values of $\boldsymbol{M}_i$ by regressing $\boldsymbol{M}_{\cdot k} = (M_{1k}, M_{2k}, \ldots, M_{nk})$, for each $k = 1, \ldots, d$, separately on the covariates $\boldsymbol{Z}$. Under H_0, $\boldsymbol{U} \sim N(\boldsymbol{0}, \boldsymbol{V})$, where $\boldsymbol{V}$ is the corresponding Fisher information matrix. Then the marginal OTU-level p-values $\boldsymbol{p} = (p_1, \ldots, p_d)$ for $\boldsymbol{\beta}$ can be obtained based on $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \ldots, \tilde{U}_d)$, where $\tilde{U}_k = U_k/V_{kk}$ and V_{kk} is the kth diagonal element of $\boldsymbol{V}$. We noted that in this paper, we choose to combine the one-sided p-values (i.e., $p^l_k = \Phi(\tilde{U}_k)$ for the lower-tail and $p^u_k = 1 - \Phi(\tilde{U}_k)$ for the upper-tail), because they account for the directionality of effects and can help boost statistical power when many OTUs have effects of the same direction. We also note that in rare situations where $V_{kk} = 0$ for some OTU k, we remove these OTUs from any subsequent analysis.

2.2 Combining P-Values from OTU-Level Tests

After getting p-values for all the OTUs (either the "extended" set or the original set), we combine them as follows. Let

$$R_k = -\log p_k, \tag{2}$$

where p_k is the p-value for testing OTU k, which can be p^l_k or p^u_k as defined above, for a particular abundance representation $\boldsymbol{M}(\boldsymbol{M}^u, \boldsymbol{M}^w, \boldsymbol{M}^{.5}$ or $\boldsymbol{M}^a)$. Since the taxa in the phylogenetic tree represent different classification levels and the abundance dispersion of different OTUs varies drastically, not all OTUs in a microbial community contribute, let alone contribute equally, to the clinical outcome of interest. Therefore, assigning different weights to OTUs according to their potential importance may enhance the statistical power of the association test. In our method, when including internal nodes, i.e., using M^u_{ik}, M^w_{ik}, or $M^{.5}_{ik}$, we use a UniFrac-like weight

$$\omega_k = \mathrm{SD}(\boldsymbol{M}_{\cdot k}) \times b_k, \qquad k = 1, \ldots, m, \tag{3}$$

where b_k is the length of the branch that leads to the kth OTU in the phylogenetic tree, and SD $(\cdot)$ stands for standard deviation. Our choice of weights takes into account both the dispersion of OTUs and their positions in the phylogenetic tree, and it is the same as that used in MiSPU and MiRKAT with UniFrac kernels if these methods are viewed as combining standardized score statistics. For M^a_{ik}, since only leaf nodes are considered, the branch length is no longer relevant; thus, we use

$$\omega_k = \mathrm{SD}(\boldsymbol{M}_{\cdot k}), \qquad k = 1, \ldots, q. \tag{4}$$

Given the weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ for all d OTUs, we can calculate

$$W_k = \omega_k R_k. \tag{5}$$

Then we sort $W_1, \ldots, W_d$ in descending order, such that $W_{(1)} \geq \cdots \geq W_{(d)}$. Let $\boldsymbol{S} = (S_1, \ldots, S_d)$ be the partial sum of $W_{(1)}, \ldots, W_{(d)}$, i.e.

$$S_k = \sum_{l=1}^{k} W_{(l)}. \tag{6}$$

For each S_k, its p-value can be defined as $P_{s_k} = \Pr(S_k \geq s_k)$, where s_k is the observed value of S_k, for $k = 1, \ldots, d$. This leads to our proposed AF statistic

$$T_{AF} = \min_{1 \leq k \leq d} P_{s_k}, \tag{7}$$

The minimizer in **Equation 7** casts some light on the associated taxa, thus, we provide a heuristic taxon selection procedure. Suppose $h = \mathrm{argmin}_{1 \leq k \leq d} P_{s_k}$, we select h taxa corresponding to the h largest W_ks as associated with the outcome. However, we caution against over-interpreting the taxon selection results, which we will further explore in **Section 3.1.2** and **Section 3.2**.

2.3 Assessing Statistical Significance by Permutation

Since the asymptotic distributions of S_k and T_{AF} are intractable when the OTU abundances are correlated, we propose to carry out the following permutation algorithm to access the null distribution of T_{AF} and estimate its corresponding p-value.

Step 1. Regress each OTU column of $\boldsymbol{M}$, $\boldsymbol{M}_{\cdot k}$, iteratively on the covariates $\boldsymbol{Z}$ to obtain the fitted OTU matrix $\hat{\boldsymbol{M}}$ and the corresponding residual matrix $\tilde{\boldsymbol{M}} = \boldsymbol{M} - \hat{\boldsymbol{M}} = \{\widetilde{M_{ij}}\}$. Calculate marginal p-values $\boldsymbol{p}$ for the OTUs according to model **Eq. 1** using $\tilde{\boldsymbol{M}}$ as $\boldsymbol{M}$. Set $\boldsymbol{p}^{(0)} = \boldsymbol{p}$.

Step 2. Permute rows of $\tilde{\boldsymbol{M}}$ for a large number of times, B, to get a set of permuted residual matrices $\{\tilde{\boldsymbol{M}}^{(1)}, \ldots, \tilde{\boldsymbol{M}}^{(B)}\}$. Obtain the permutation set of p-values $\{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(B)}\}$, by refitting the regression model with the permuted residuals for $b = 1, \ldots, B$.

Step 3. Follow **Equations 2–6** to obtain $\boldsymbol{S}^{(b)} = (S^{(b)}_1, \ldots, S^{(b)}_d)$, for $b = 0, 1, \ldots, B$, where $\boldsymbol{S}^{(0)}$, corresponding to $\boldsymbol{p}^{(0)}$, denoting the statistic based on the original data.

Step 4. For each $b = 0, 1, \ldots, B$ and $k = 1, \ldots, d$, calculate

$$P_{S_k}^{(b)} \approx \frac{1}{B+1} \sum_{b^*=0}^{B} I\left\{S_k^{(b^*)} \geq S_k^{(b)}\right\}.$$

Then we can get the observed and permuted AF statistics $T_{AF}^{(b)} = \min_{1 \leq k \leq d} P_{S_k}^{(b)}$, for $b = 0, 1, \ldots, B$.

Step 5. The p-value $P_{AF}^{(b)}$ of the AF statistic $T_{AF}^{(b)}$ can be approximated by

$$P_{AF}^{(b)} = \Pr\left\{T_{AF} \leq T_{AF}^{(b)}\right\} \approx \frac{1}{B+1} \sum_{b^*=0}^{B} 1\left\{T_{AF}^{(b^*)} \leq T_{AF}^{(b)}\right\},$$

where $b = 0, 1, \ldots, B$.

Note that in Step 1 and 2, we permute the residuals of regression $\boldsymbol{M}_{\cdot k}$ on $\boldsymbol{Z}$ and fit a generalized linear model using the permuted residuals, which preserves the correlation among covariates $\boldsymbol{Z}$ and abundance representation $\boldsymbol{M}$ even after permutation. We also noted that we used index $b = 0$ to denote the statistics calculated from the original observed data. Therefore $P_{AF}^{(0)}$ is the final p-value of our proposed AF statistic if there is only one list of OTU-level p-values $\boldsymbol{p} = (p_1, \ldots, p_d)$ to combine, e.g., only $\boldsymbol{p}^l$ using $\boldsymbol{M}^u$. Besides, we also calculated $P_{AF}^{(b)}$ for $b = 1, \ldots, B$, which are B permutations of the AF p-value. These permutations can be further used to combine the results of multiple AF p-values generated by combining p-values from our multiple OTU abundance representations, which we discuss next.

2.4 Combining Multiple AF Tests

In the method described in the previous subsections, there are multiple variations or factors that can affect the performance of the test under different scenarios, including the choice of OTU-level tests, the transformation from relative abundance, $\boldsymbol{A}$, to an abundance representation, $\boldsymbol{M}$, the usage of one-sided or two-sided p-values, and the weights used in the combination step. Therefore, to construct a statistical test that is robust under various scenarios, it is often desirable to combine the results from multiple tests based on different parameter choices. We therefore, propose to combine the results of multiple AF tests with different parameter selections to form a unified test.

The p-value combination approach that we described previously in **Section 2.2** can be viewed as a general method for combining multiple p-values with or without weights, as long as we can obtain a permuted sample while preserving the correlation among them. We define operation $AF\{\boldsymbol{p}[;\ \boldsymbol{\omega}]\}$ as the procedure that combined a p-value vector $\boldsymbol{p}$ with optional weight vector $\boldsymbol{\omega}$, which defaults to ones when omitted. By using this AF operator, we can redefine our MiAF method that combines results from different choices of OTU-level test p-values, weights, and abundance representations. For illustration purpose, in the rest of our paper, we combine results from lower- and upper-tail p-values using the unweighted ($\boldsymbol{M}^u$), weighted ($\boldsymbol{M}^w$), square-root ($\boldsymbol{M}^{\cdot 5}$) abundance representations for "extended" OTUs and their corresponding weights as defined above, as well as the abundance representations for leaf nodes only ($\boldsymbol{M}^a$) and its corresponding weights. Specifically, $\boldsymbol{p}^{ul}$ and $\boldsymbol{p}^{uu}$ denote the lower- and upper-tail p-values of the OTU-level tests using $\boldsymbol{M}^u$. Similarly, we use $\boldsymbol{p}^{wl}$ and $\boldsymbol{p}^{wu}$ for $\boldsymbol{M}^w$, $\boldsymbol{p}^{\cdot 5l}$ and $\boldsymbol{p}^{\cdot 5u}$ for $\boldsymbol{M}^{\cdot 5}$, and $\boldsymbol{p}^{al}$ and $\boldsymbol{p}^{au}$ for $\boldsymbol{M}^a$.

With the associated weights denoted as $\boldsymbol{\omega}^a$, $\boldsymbol{\omega}^w$, $\boldsymbol{\omega}^{\cdot 5}$ and $\boldsymbol{\omega}^a$, respectively, we can obtain the p-value for each of the eight community-level MiAF tests by combining the corresponding OTU-level p-value vectors and the corresponding weight vectors using the AF operator defined above; details are given in the 5^{th} and 6^{th} columns of **Table 1**. The two one-sided community-level tests are then combined to form a two-sided test, again using the AF operator, for each of the four abundance measure tests (column 7 of **Table 1**). Our eventual test statistic, MiAF, combines the unweighted UniFrac-like test p-value P_{MiAF_u}, the weighted UniFrac-like test p-value P_{MiAF_w}, the generalized UniFrac-like test p-value $P_{MiAF_{.5}}$ and the leaf-nodes-only test p-value P_{MiAF_a}, again using AF operator (last row of **Table 1**). We declare that the microbial community is significantly associated with the clinical outcome if P_{MiAF} is smaller than a prespecified significance level α.

3 RESULTS

3.1 Simulation Study

3.1.1 Simulation Strategy

We conducted simulation studies to investigate whether MiAF correctly controls type I error and to evaluate the performance of MiAF in a wide range of scenarios. We generated unobvserved absolute abundances and read counts of OTUs using the R package SparseDOSSA2 which can parameterize real microbial profiles and then simulate new profiles based on the estimated parameters (Ma et al., 2021). SparseDOSSA2 depicts the unobserved absolute abundance *via* a Gaussian copula model with zero-inflated log normal marginal distributions. To address the identifiability issue, it imposes L_1 penalization on the correlation matrix. Using SparseDOSSA2 package, we first parameterized a real upper-respiratory-tract microbiome data set consisting of 856 OTUs and 60 samples (Charlson et al., 2010). The penalizing tuning parameter was chosen to be 0.1 since it achieved the largest likelihood among $\{0.1, 0.2, \ldots, 1\}$. 616 OTUs remained after discarding the OTUs with only one non-zero count across the samples. Then the microbial community profiles for 616 OTUs and 100 samples, including unobserved absolute abundance and read counts, based on the estimated parameters were simulated. We denoted the simulated absolute abundance matrix by $\boldsymbol{X}$, where X_{ij} was the absolute abundance of OTU j in sample i.

To evaluate our method, we implemented three simulation scenarios where the OTUs were divided into different clusters and related to both binary and continuous outcomes in different ways. The clustering on OTUs was based on partitioning around medoids (Kaufman and Rousseeuw, 1990) with cophenetic distance (Sokal and Rohlf, 1962). We chose three cluster numbers: 10, 22 and 29, corresponding to the first three local maxima of the mean silhouette values shown in **Supplementary Figure S1** of Supplementary Material. Under scenario 1, the 616 OTUs were grouped into 22 clusters. The abundance varied greatly among these 22 clusters. In order to test our new

TABLE 1 | MiAF implementation algorithm.

Tests	Abundance measure	Relationship to A	Phylogenetic information	Single measure		Combine multiple measures
				Lower-Tail	Upper-Tail	
MiAF$_u$	$\boldsymbol{M}^u$	$M^u_{ik} = I(A_{ik} > 0)^*$	✓	$P^{ul} = AF\{\boldsymbol{p}^{ul}; \boldsymbol{\omega}^u\}$	$P^{uu} = AF\{\boldsymbol{p}^{uu}; \boldsymbol{\omega}^u\}$	$P_{MiAF_u} = AF\{(P^{ul}, P^{uu})^T\}$
MiAF$_w$	$\boldsymbol{M}^w$	$M^w_{ik} = A_{ik}{}^*$	✓	$P^{wl} = AF\{\boldsymbol{p}^{wl}; \boldsymbol{\omega}^w\}$	$P^{wu} = AF\{\boldsymbol{p}^{wu}; \boldsymbol{\omega}^w\}$	$P_{MiAF_w} = AF\{(P^{wl}, P^{wu})^T\}$
MiAF$_{.5}$	$\boldsymbol{M}^{.5}$	$M^{.5}_{ik} = A_{ik}{}^{.5\star}$	✓	$P^{.5l} = AF\{\boldsymbol{p}^{.5l}; \boldsymbol{\omega}^{.5}\}$	$P^{.5u} = AF\{\boldsymbol{p}^{.5u}; \boldsymbol{\omega}^{.5}\}$	$P_{MiAF_{.5}} = AF\{(P^{.5l}, P^{.5u})^T\}$
MiAF$_a$	$\boldsymbol{M}^a$	$M^a_{ik} = A_{ik}{}^\dagger$	✗	$P^{al} = AF\{\boldsymbol{p}^{al}; \boldsymbol{\omega}^a\}$	$P^{au} = AF\{\boldsymbol{p}^{au}; \boldsymbol{\omega}^a\}$	$P_{MiAF_a} = AF\{(P^{al}, P^{au})^T\}$
MiAF	—	—	—	—	—	$P_{MiAF} = AF\{(P_{MiAF_u}, P_{MiAF_w}, P_{MiAF_{.5}}, P_{MiAF_a})^T\}$

*k = 1, . . ., m.
†k = 1, . . ., q.

method in broader circumstances, we performed the simulation analysis assuming that the outcome is truly associated with each cluster of OTUs iteratively instead of evaluating the performance on only a few clusters. The binary outcome Y_i for sample i, i = 1, . . ., 100, was simulated based on model

$$\text{logit}\left(E\left(Y_i \mid X_i, \boldsymbol{Z}_i\right)\right) = 0.5\,\text{scale}\left(Z_{i1} + Z_{i2}\right) + \beta\,\text{scale}\left(\sum_{j\in C} X_{ij}\right). \tag{8}$$

We simulated continuous outcomes under the model

$$Y_i = 0.5\,\text{scale}\left(Z_{i1} + Z_{i2}\right) + \beta\,\text{scale}\left(\sum_{j\in C} X_{ij}\right) + \epsilon_i, \tag{9}$$

where $\epsilon_i \sim N(0, 1)$. For both binary and continuous outcomes, Z_{i1} and Z_{i2} were covariates, and C was the set of OTUs that belong to a selected cluster. The scale(·) function standardizes the sample mean to 0 and standard deviation to 1. Z_{i1} was drawn from a Bernoulli distribution with success probability 0.5 independently. For Z_{i2}, we consider two situations where Z_{i2} and the abundance of the microbial community $\boldsymbol{X}_{i\cdot}$ are either independent or correlated. In the independent case, Z_{i2} was generated from standard normal distribution N (0, 1), and the effect size β was set as 0.6, 0.8, 1.2, 1.6 and 2 for binary outcomes, and 0.2, 0.4, 0.6, 0.8 and 1 for continuous outcomes to mimic different levels of association strength between the OTUs and the clinical outcome. In the correlated case, we let $Z_{i2} = \text{scale}(\sum_{j\in C} X_{ij}) + \tau$, where $\tau \sim N(0, 1)$ and the effect size β was set to be twice as large as the corresponding value in the independent case, in order to show a clearer difference among the methods compared.

Under scenario 2, we divided the 616 OTUs into 10 clusters and simulated the data on all clusters following the same settings. For scenario 3, all OTUs were divided into 29 clusters following the same procedure.

Under all three simulation scenarios, the performance of MiAF was compared to MiRKAT, aMiSPU, OMiAT, aMiAD and MiHC. We did not include PERMANOVA because it is essentially equivalent to MiRKAT without covariates (Zhao et al., 2015). aMiSPU combines unweighted and weighted UniFrac versions of test. MiRKAT combines four kernels, including the unweighted and weighted UniFrac, a generalized UniFrac with tuning parameter at 0.5, and the Bray-Curtis. MiAF combines the unweighted, weighted, generalized UniFrac-like and the leaf-nodes-only test p-values. We used default setting of OMiAT, which includes all the kernels in MiRKAT but with an addition of the Jaccard distance. aMiAD combines six alpha diversity metrics as its default setting, which includes Richness, Shannon, Simpson, phylogenetic diversity (PD), phylogenetic entropy (PE) (Allen et al., 2009) and phylogenetic quadratic entropy (PQE) (Rao, 1982). MiHC combines the unweighted higher criticism test, weighted higher criticism test and Simes test, and the candidate set for both higher criticism tests to modulate low sparsity level was set as {1, 3, 5, 7, 9}. We set the significance level to be 0.05 for each test. When evaluating the type I error under all the simulation scenarios, we simulated data according to model (**Eqs 8** and **9**) by setting $\beta = 0$. We set the number of permutation for all five methods as 10,000 to assess their ability for correct control of type I error. When comparing power, the number of permutation was set to be 1,000. All simulation results were based on 1,000 independent replicates.

We investigated the performance of the proposed heuristic taxa selection procedure when setting $\beta > 0$ in the model (**Eqs 8** and **9**). Although only tip nodes were explicitly assumed to be associated with the outcome in the simulation setting, we also viewed the internal nodes as associated taxa if any of their descendants was associated. Since the outcome were generated to be positively correlated with the abundances of OTUs within the associated cluster, we recorded the number of every taxon being selected from upper-tail p-values in the 1,000 independent replicates.

3.1.2 Simulation Results

Figure 2 shows the statistical power for binary outcomes under scenario 1, where 616 OTUs were partitioned into 22 clusters, and when both covariates Z_{i1} and Z_{i2} are independent of the microbial community. The cluster size and mean absolute abundance varies greatly among 22 clusters (see details in **Supplementary Table S1B** of Supplementary Material), covering different underlying association patterns. We evaluated the performance of all the methods under situations where each phylogenetic cluster of OTUs was set to be associated with the binary outcome successively. The power of the six methods was plotted against clusters sorted by the sum of estimated mean absolute abundance of OTUs within the cluster that was truly associated from the greatest to the least, representing the total strength of signals. As expected, for each associated cluster community, the statistical power increased as the effect size β increased. For aMiSPU,

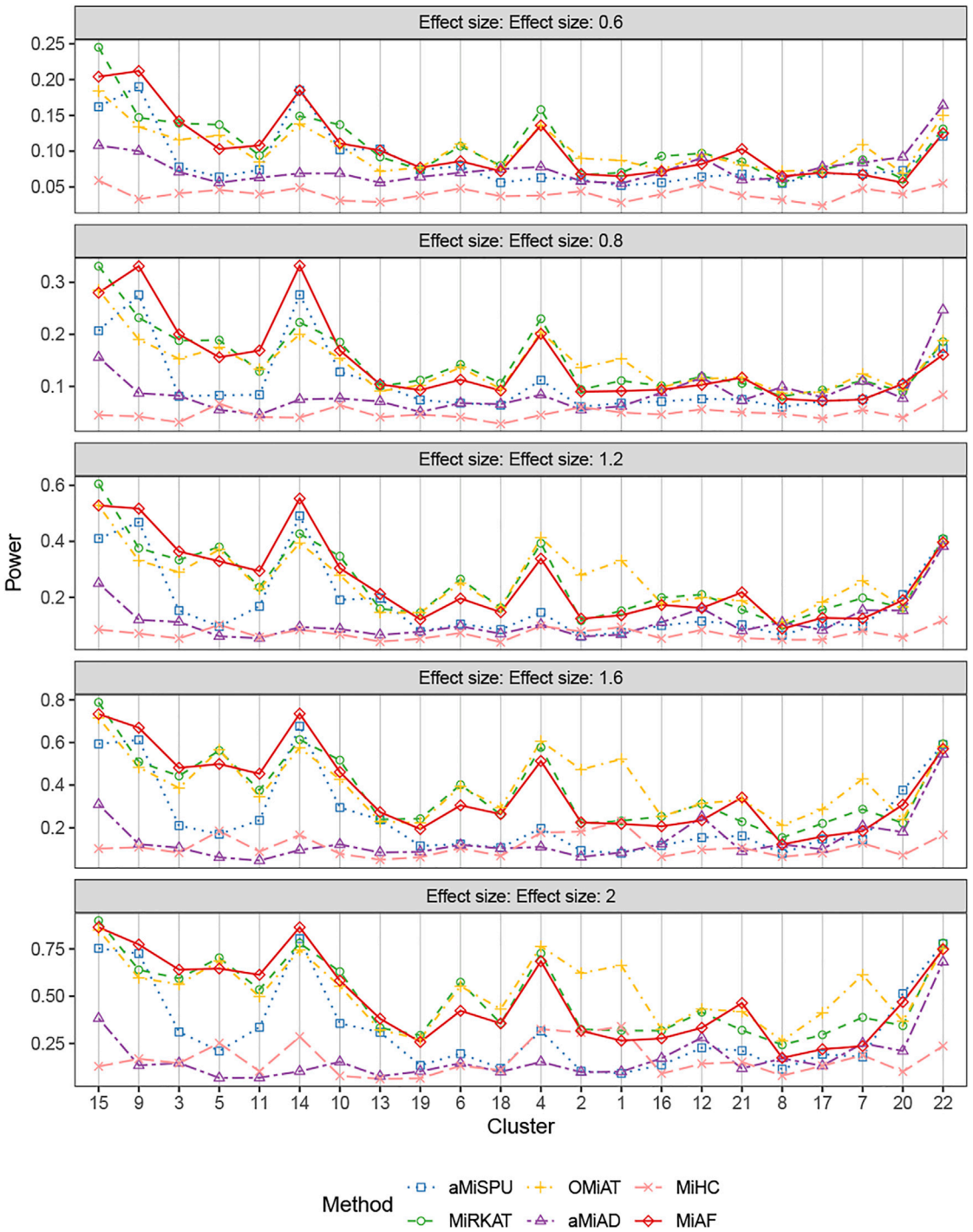


FIGURE 2 | Power comparison for binary outcomes under the independent case of scenario 1. A total of 616 OTUs were divided into 22 clusters. The covariates Z_{i2} and OTUs $\boldsymbol{X}_i$ were independent. The effect size was set as 0.6, 0.8, 1.2, 1.6 and 2. The 22 clusters were sorted by the sum of estimated mean absolute abundance of the OTUs within the cluster that was truly associated from the greatest to the least.

MiRKAT and MiAF, the performance of the unweighted version of tests was outperformed by the weighted version of tests in the majority of the clusters, with exceptions of clusters 7, 12, 16, 20 and 22. Another observation for all methods was that the combined tests lose only a little power compared to the best one of their corresponding component tests, which justifies the use of a combined or optimal test to draw a unified conclusions from multiple parameter choices. Therefore, we focused on

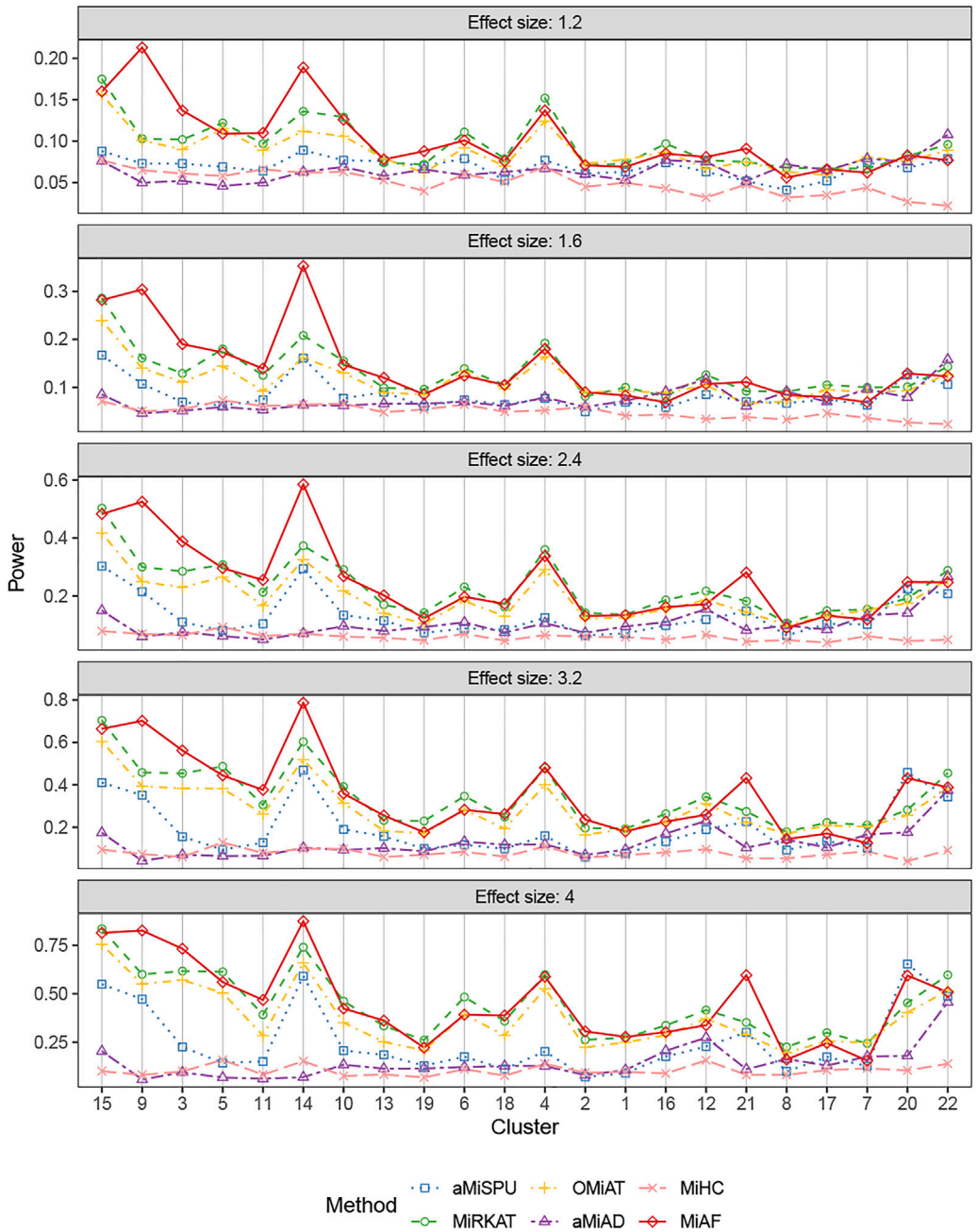


FIGURE 3 | Power comparison for binary outcomes under the correlated case of scenario 1. A total of 616 OTUs were divided into 22 clusters. The covariates Z_{i2} and OTUs $\boldsymbol{X}_i$ were correlated. The effect size was set as 1.2, 1.6, 2.4, 3.2 and 4. The 22 clusters were sorted by the sum of estimated mean absolute abundance of the OTUs within the cluster that was truly associated from the greatest to the least.

comparing the performance of the combined version of the six tests in the rest of this section. When the total sum of estimated mean absolute abundance of OTUs within the associated clusters was relatively large (around top 40% of the sum of absolute abundance of associated OTUs among the 22 clusters), MiAF either outperformed the other five methods or was commensurate with the best of the other five. When MiAF was not the best, either MiRKAT or OMiAT was always among the top, where the power of OMiAT was predominantly driven by MiRKAT. When the sum of estimated mean absolute abundance of the associated OTUs was relatively small (around lower 60% among the 22 clusters), OMiAT had overall the best performance. In most cases, MiAF outperformed the inferior methods by a large margin even if it was not the best.

The results for binary outcomes under scenario 1 with covariate Z_{i2} correlated with the OTU abundance were

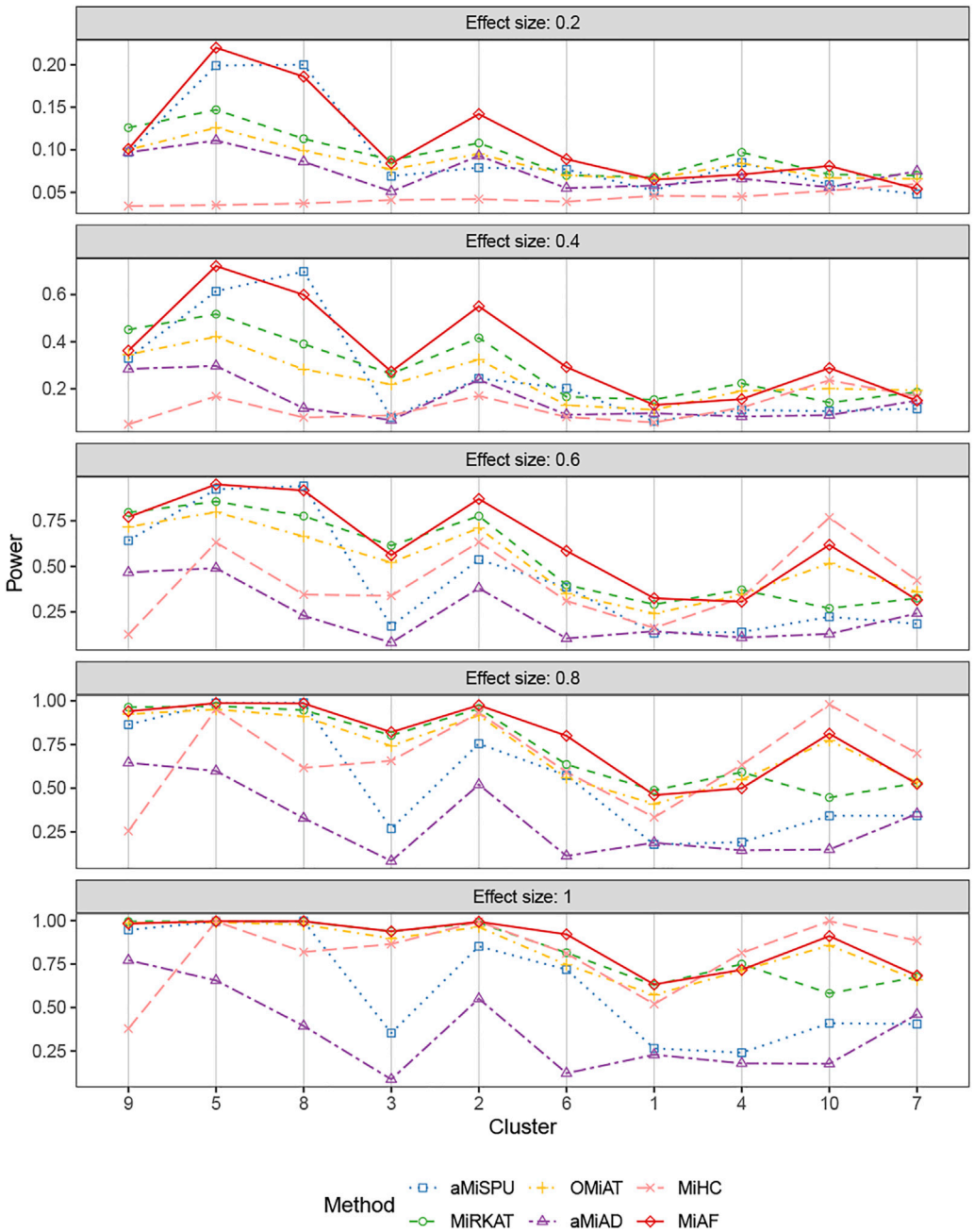


FIGURE 4 | Power comparison for continuous outcomes under the independent case of scenario 2. A total of 616 OTUs were divided into 10 clusters. The covariates Z_{i2} and OTUs $\boldsymbol{X}_i$ were independent. The effect size was set as 0.2, 0.4, 0.6, 0.8 and 1. The 10 clusters were sorted by the sum of estimated mean absolute abundance of the OTUs within the cluster that was truly associated from the greatest to the least.

shown in **Figure 3**. Similar to the independent covariate case, the weighted tests possessed relatively higher statistical power with exceptions of clusters 7, 12, 16, 20, 21 and 22. In terms of the combined test, the advantage of MiAF over the other methods was more prominent than that in the independent case. In all the clusters except for cluster 7 and 8 where several methods were on par, MiAF achieved a dominant position over the other five methods or was a close second. We observed distinct advantage of MiAF in clusters 3, 9, 14 and 21, where MiAF had moderate power even when the effect size was small. It was interesting to see that the unweighted tests achieved their greatest power in cluster 22 where the mean OTU abundance was the lowest among all clusters. It confirmed that the unweighted tests are more powerful when clinical outcomes are associated with rare microbial taxa (Chen et al., 2012).

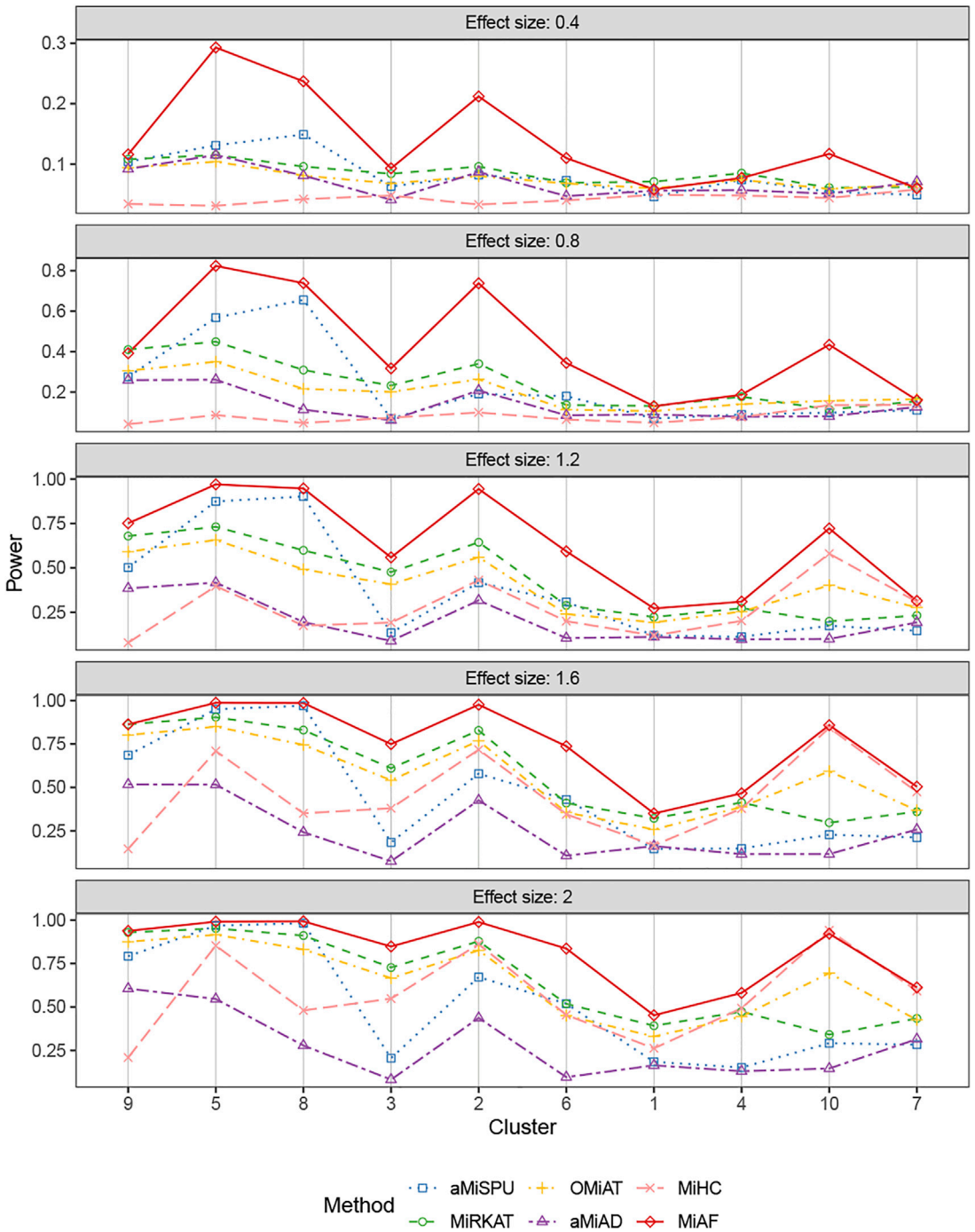


FIGURE 5 | Power comparison for continuous outcomes under the correlated case of scenario 2. A total of 616 OTUs were divided into 10 clusters. The covariates Z_{i2} and OTUs $\boldsymbol{X}_i$ were correlated. The effect size was set as 0.4, 0.8, 1.2, 1.6 and 2. The 10 clusters were sorted by the sum of estimated mean absolute abundance of the OTUs within the cluster that was truly associated from the greatest to the least.

The simulation results for binary outcomes under scenario 2 and scenario 3 showed similar results, when the OTUs were partitioned into 10 or 29 clusters respectively. The power comparisons were shown in **Supplementary Figures S2–S5**. Our method, MiAF, achieved a dominant position over other methods consistently in correlated cases, where the existence of correlation between microbes and covariates is more biologically relevant in practice. MiAF performed equivalently well with OMiAT and MiRKAT in the independent cases when the sum of absolute abundances of the associated OTUs was relatively

TABLE 2 | Type I error rates under independent case and mean type I error rates under correlated cases for both binary and continuous outcomes.

Simulation scenarios		aMiSPU	MiRKAT	OMiAT	aMiAD	MiHC	MiAF
Binary response	Independent case	0.059	0.048	0.047	0.050	0.025	0.055
	Correlated case, 10 clusters	0.046	0.042	0.046	0.043	0.032	0.051
	Correlated case, 22 clusters	0.048	0.047	0.049	0.047	0.032	0.048
	Correlated case, 29 clusters	0.050	0.046	0.047	0.048	0.025	0.048
Continuous response	Independent case	0.049	0.047	0.059	0.055	0.038	0.050
	Correlated case, 10 clusters	0.043	0.045	0.044	0.045	0.047	0.042
	Correlated case, 22 clusters	0.047	0.048	0.047	0.046	0.039	0.049
	Correlated case, 29 clusters	0.050	0.045	0.049	0.048	0.032	0.047

large (around top 70% and 30% among the 10 and 29 clusters respectively); while when the sum of absolute abundance of the associated OTUs was relatively small around lower 30% and 70% among the 10 and 29 clusters respectively, OMiAT was always in the lead. The power of MiAF was affected by the effect size and underlying association patterns, but not the direction of the effect (see the power comparison under the independent case of 10 clusters with positive and negative effects in **Supplementary Figure S6**).

Figures 4, **5** displays the statistical power for continuous outcomes under scenario 2 for independent and correlated cases respectively, where 616 OTUs were divided into 10 clusters. The largest cluster consists of 171 OTUs (27.76%), and the sizes of the rest clusters are between 23 (3.41%) and 68 (11.04%) (see details in **Supplementary Table S1A**). In contrast to clearly different power comparison trend between independent and correlated cases for binary outcomes, the comparative power among the six methods was mainly affected by the associated clusters for continuous outcomes. MiAF continued to thrive when the sum of estimated mean absolute abundance of the OTUs within the selected cluster was relatively large (top 70% among 10 clusters for independent case and all 10 clusters for correlated case). We observed a great disparity in the performance of MiHC between binary and continuous outcomes, where MiHC was more capable of detecting the association between microbial communities and a continuous outcome. Besides, MiHC was barely able to detect the association for small effect size scenarios, and its power surged when the effect size raised to high level. MiHC had the greatest power among all the methods for relatively small sum of absolute abundance of the associated OTUs (around lower 2/3 and 1/2 for independent and correlated case respectively among 22 and 29 clusters), especially in some results with 22 or 29 clusters where the associated clusters tended to be in small size due to the large number of clustering (see **Supplementary Figures S7–S10**).

Empirical Type I error rates of the six methods across different simulation scenarios are shown in **Table 2**. Under the null model of independent case where the selected OTU cluster did not play a role, we had one unified assessment of type I error. For the correlated case, we averaged the type I error rates over all clusters within each scenario. The details of type I error rates for each cluster are provided in **Supplementary Tables S2–S7** for binary and continuous responses respectively. Further, we investigated the Type I error rates for the independent case with QQ-plot of p-values in $-\log_{10}$ scale against a uniform distribution between 0 and 1 shown in **Supplementary Figures S11, S12**. We can see that the error rate was conservative for MiHC under binary responses, and that it was well under control for other methods (~ 0.05) in general, which confirmed that our method is statistically valid.

We compared the taxon selection results with the truth in our simulation settings. To demonstrate the performance of our heuristic taxon selection procedure, we took cluster 1 out of 10 clusters for continuous outcomes under independent case with effect size 1 as an example shown in **Supplementary Figure S13** (see more results in **Supplementary Figures S14–S16**). The most often selected taxa over 1,000 replicates tended to be in high abundance, belonging to the truly associated cluster. MiAF had more difficulties in identifying associated taxa with low abundance, since the selection of low abundance taxa suffered from random noise, which renders the selection results of low abundance taxa unstable and unreliable. Therefore, the taxon selection result was more useful for abundant taxa, leading to more trustworthy insight into selecting taxa at relatively higher level of the phylogenetic tree in general, as their counts were aggregated from their descendants. To help navigate the taxa selection result and focus on abundant taxa only, we provide a visualization tool in our R package where the transparency of each branch was set according to the abundance of its node. The tendency to discover abundant associated taxa was consistent with the prominent performance of MiAF when the sum of absolute abundance of associated OTUs was large in the previous power results. We called the 10% most often selected taxa over 1,000 replicates as selected taxa in a simulation scenario, or otherwise as non-selected taxa to err on the conservative side. Under the independent case for continuous outcomes with effect size 1 where 616 OTUs were divided into 10 clusters, we also provided the sensitivity and specificity for abundant taxa, specifically taxa with abundance over 75%, 80% and 85% quantiles respectively in **Supplementary Table S8**. The overall specificity was considerably high, although the sensitivity was lower. It suggests that a subset of the associated taxa can be identified, and that we are unlikely to select wrong taxa based on our heuristic taxon selection algorithm. It should be noted that the taxa selection result is only exploratory and should not be over-interpreted.

TABLE 3 | P-values of aMiSPU, MiRKAT, OMiAT, aMiAD, MiHC and MiAF for the association test between smoking status and throat microbial community.

	aMiSPU	MiRKAT	OMiAT	aMiAD	MiHC	MiAF
p-value	0.0025	0.0046	0.0096	0.0167	0.2249	0.0025

TABLE 4 | P-values of aMiSPU, MiRKAT, OMiAT, aMiAD, MiHC and MiAF for the association test between HIV infectious status and gut microbial community.

	aMiSPU	MiRKAT	OMiAT	aMiAD	MiHC	MiAF
p-value	0.0114	0.0002	0.0001	0.0002	0.0001	0.0003

3.2 Real Data Analysis

3.2.1 Application to a Throat Microbiome Dataset

In our first real data application to demonstrate the utility of our proposed MiAF, we applied it and the competing methods to a profiling study of microbial communities in the upper respiratory tract to explore the effect of cigarette smoking (Charlson et al., 2010). In the study, microbiota were collected from the right and left nasopharynx and oropharynx of 29 smokers and 33 healthy non-smokers. After PCR amplification and QIIME pipeline, OTUs were constructed at 97% similarity. The preprocessed dataset is included in many statistical software packages such as GUniFrac (Chen et al., 2012), MiRKAT (Zhao et al., 2015) and MiSPU (Pan et al., 2014) as the testing data, which contain information on 856 OTUs in 60 samples (28 smokers and 32 nonsmokers), a slightly reduced data set from the original study. Our application used this dataset following the papers of MiRKAT and aMiSPU.

We applied MiRKAT, aMiSPU, OMiAT, aMiAD, MiHC and MiAF on this dataset to test the association between smoking and microbial community composition while controlling for gender. **Table 3** presents p-values of these six methods. The combined MiAF generated a p-value of 0.0025, which confirmed the results published in previous studies that the association between the microbial community and smoking status remained significant while adjusting for possible confounders (Brook and Gober, 2008; Charlson et al., 2010; Schenck et al., 2016). MiHC was the only method that failed to detect such association among the six methods. The unweighted test of aMiSPU and MiAF_u, as well as aMiAD using alpha diversity metrics Richness, Shannon, Simpson and phylogenetic diversity, alone failed to detect such association at significance level 0.05, although their corresponding combined results were significant. All the component tests of MiHC failed to detect any association in this dataset (see results of all the component tests of the six methods in **Supplementary Table S9**).

Besides an overall evaluation of association, selecting associated taxa in a microbial community is also of interest. MiAF provides a heuristic taxon selection by choosing the top h taxa in the p-value combination step, where h is the minimizer of **Equation 7**. **Supplementary Figure S17** shows the selected associated taxa for this throat microbiome dataset. The phylogenetic tree was plotted using the R package ggtree (Yu et al., 2018). MiAF detected 1 associated node to be under-presented in the smokers based on lower-tail p-values, and it detected 128 associated nodes to be over-presented from upper-tail p-values as well.

3.2.2 Application to a Stool Microbiome Dataset

HIV infection induces substantial gut microbiome alterations. Lozupone et al. (Lozupone et al., 2013) revealed that HIV infection was associated with highly characteristic gut microbial community changes through 16 S rRNA sequencing of feces. In our second real data application, we downloaded the processed OTU data consisting of 10104 100% OTUs, i.e., ASVs, from the MicrobiomeHD database (Duvallet et al., 2017). After matching the samples to their clinical data, our analysis was conducted based on 22 HIV-infected individuals and 13 HIV-negative controls. After excluding OTUs with all zero counts in the 35 samples, 9,460 OTUs remained in the analysis. We built the phylogenetic tree using the QIIME2 pipeline (Bolyen et al., 2019).

We investigated the association between disease status and the overall microbial community composition using the six methods all based on 10,000 permutations, adjusting for potential confounder age. **Table 4** shows the p-values generated by the six methods, where all the methods were able to detect the association at significance level 0.01 except for aMiSPU. While the unweighted test of aMiSPU and aMiAD using Shannon, Simpson and phylogenetic diversity, as well as the Simes test combined by MiHC failed to detect any association, the results of all the other component tests were significant at the 0.05 level (see details in **Supplementary Table S10**). As in the first application, we were also interested in finding individual taxa that are thought to be associated with HIV status. To this end, MiAF detected 224 and 57 associated nodes from under- and over-presented in the HIV-infected individuals respectively (phylogenetic tree plot was not included because it was hardly readable due to the large number of OTUs).

4 DISCUSSION

In this paper, we proposed an adaptive p-value combination approach to construct a community-level association test from those that are OTU-level based. In general, combining OTU-level tests without adaptation or weighting may not generate comparable statistical power to sophisticated methods specifically designed for community-level association test. To demonstrate the usage and statistical power of the proposed approach, we constructed a community-level test, MiAF, by combining the p-values of univariate score tests using UniFrac-like and Bray-Curtis-like transformations and weighting scheme, and showed that its statistical power is comparable or better than methods specifically designed for community test. We chose to combine the p-values of score statistics to make it a fair comparison to the competing methods, because the performance of our method depends on the selection of univariate tests and the aMiSPU, MiRKAT and OMiAT test statistics can all be viewed as functions of the score statistics with similar weight selection.

It should be noted that the aMiSPU test can also be viewed as a test that combines OTU-level score test statistics. However, comparing to score statistic, *p*-value is a much more readily available statistic for various univariate testing methods. Although we demonstrated the usage of our proposed method using score tests, *p*-values are the quantities that we ultimately combine. This leads to the flexibility to our proposed framework since other tests can be combined into community-level tests, as long as they satisfy two conditions: 1) *p*-values (ideally one-sided) are available and can correctly control type I error; and 2) permutation or resampling methods exist to generate a reference distribution for the *p*-values while maintaining the correlation structure among the OTUs. We note that these two condition are met by a lot of tests, such as the tests in various regression models, where we can adopt a similar permutation procedure that permutes the residual of the condition of interest that regressed on the confounding covariates. For example, using this strategy, it will be relatively easy to construct a community-level test for survival outcome by combining any survival models, such as the Cox model or the accelerated failure time model. In addition, it is also possible to combine OTU-level tests to accommodate longitudinal outcomes or longitudinal microbiome measurements, which is our next topic in the future research.

A side product of our method is taxon selection, which is naturally provided by the minimizer. By plotting the selected taxa along the phylogenetic tree, we can see that they tend to occupy consecutive branches that leads to much fewer OTUs, which matches our intuition, because if a species is over-presented, the taxa in the upper hierarchy (such as genus, family, order, etc.) that contains the species should also be over-presented. However, this variable selection is only heuristic and is not the focus of this paper, because the *p*-values we combined are from univariate models, which perform marginal tests not conditional on other OTUs. Therefore, the OTUs selected are only marginally related to the outcome. It is still possible that some of the selected OTUs correlate to the outcome through other OTUs, which is a limitation of the proposed method. Another limitation is the relatively slow computational speed compared to aMiSPU, MiRKAT, OMiAT and MiHC when there are a large number of taxa, but it is faster than aMiAD. When analyzing the data set of our first real data application in a laptop with 8-core CPU and 8 GB unified memory, it takes 7 s for aMiSPU, 3 s for MiRKAT, 17 s for OMiAT, 21 s for MiHC, 5 min 45 s for MiAF, and 7 min 41 s for aMiAD. Despite relatively slower computational speed, it is still computationally feasible to apply MiAF to real data sets given that MiAF will only need to be performed once on the data set to test the association. Improving the computational speed of our method is one of our future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The throat microbiome dataset is available in R package MiSPU, and the stool microbiome dataset (hiv_lozupone) is openly available in MicrobiomeHD at https://zenodo.org/record/1146764#.Ylh_1i1h1PM.

AUTHOR CONTRIBUTIONS

CS developed the new method. QC implemented the methods and performed the simulation study and real data analysis. SL supervised the project. SL, CS, and QC wrote the manuscript.

FUNDING

Part of CS's time is supported by NSF grant 1921592 and NCATS grant 1UL1TR002733.

ACKNOWLEDGMENTS

The authors thank Ohio Supercomputer Center (Center, 1987) for providing the computing resource for the simulation and real data application. The authors thank the reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.846258/full#supplementary-material

REFERENCES

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., et al. (2013). Human Gut Microbiome and Risk for Colorectal Cancer. *J. Natl. Cancer Inst.* 105, 1907–1911. doi:10.1093/jnci/djt300

Allen, B., Kon, M., and Bar-Yam, Y. (2009). A New Phylogenetic Diversity Measure Generalizing the shannon index and its Application to Phyllostomid Bats. *The Am. Naturalist* 174, 236–243. doi:10.1086/600101

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems* 2. doi:10.1128/mSystems.00191-16

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using Qiime 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9

Bray, J. R., and Curtis, J. T. (1957). An Ordination of the upland forest Communities of Southern wisconsin. *Ecol. Monogr.* 27, 326–349. doi:10.2307/1942268

Brook, I., and Gober, A. E. (2008). Recovery of Potential Pathogens in the Nasopharynx of Healthy and Otitis Media-Prone Children and Their Smoking and Nonsmoking Parents. *Ann. Otol Rhinol Laryngol.* 117, 727–730. doi:10.1177/000348940811701003

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, 581–583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME Allows Analysis of High-Throughput

Community Sequencing Data. *Nat. Methods* 7, 335–336. doi:10.1038/nmeth.f.303

[Dataset] Center, O. S. (1987). *Ohio Supercomputer center*. Columbus, Ohio, USA: Ohio supercomputer center.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PloS one* 5, e15216. doi:10.1371/journal.pone.0015216

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating Microbiome Composition with Environmental Covariates Using Generalized Unifrac Distances. *Bioinformatics* 28, 2106–2113. doi:10.1093/bioinformatics/bts342

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of Gut Microbiome Studies Identifies Disease-specific and Shared Responses. *Nat. Commun.* 8, 1784. doi:10.1038/s41467-017-01973-8

Faith, D. P. (1992). Conservation Evaluation and Phylogenetic Diversity. *Biol. conservation* 61, 1–10. doi:10.1016/0006-3207(92)91201-3

Hasan, N. A., Young, B. A., Minard-Smith, A. T., Saeed, K., Li, H., Heizer, E. M., et al. (2014). Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. *PLoS One* 9, e97699. doi:10.1371/journal.pone.0097699

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone.1. *New Phytol.* 11, 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Kaufman, L., and Rousseeuw, P. J. (1990). Partitioning Around Medoids (Program Pam). *Finding groups Data introduction cluster Anal.* 344, 68–125.

Koh, H. (2018). An Adaptive Microbiome α-diversity-based Association Analysis Method. *Sci. Rep.* 8, 18026. doi:10.1038/s41598-018-36355-7

Koh, H., Blaser, M. J., and Li, H. (2017). A Powerful Microbiome-Based Association Test and a Microbial Taxa Discovery Framework for Comprehensive Association Mapping. *Microbiome* 5, 45–15. doi:10.1186/s40168-017-0262-x

Koh, H., and Zhao, N. (2020). A Powerful Microbial Group Association Test Based on the Higher Criticism Analysis for Sparse Microbial Association Signals. *Microbiome* 8, 63–16. doi:10.1186/s40168-020-00834-9

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. doi:10.1053/j.gastro.2014.02.009

Lasken, R. S. (2012). Genomic Sequencing of Uncultured Microorganisms from Single Cells. *Nat. Rev. Microbiol.* 10, 631–640. doi:10.1038/nrmicro2857

Ley, R. E. (2010). Obesity and the Human Microbiome. *Curr. Opin. Gastroenterol.* 26, 5–11. doi:10.1097/mog.0b013e328333d751

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors that Structure Microbial Communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi:10.1128/aem.01996-06

Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., et al. (2013). Alterations in the Gut Microbiota Associated with Hiv-1 Infection. *Cell host & microbe* 14, 329–339. doi:10.1016/j.chom.2013.08.006

Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., et al. (2021). A Statistical Model for Describing and Simulating Microbial Community Profiles. *Plos Comput. Biol.* 17(9), e1008913. Public Library of Science San Francisco, CA, USA.

McArdle, B. H., and Anderson, M. J. (2001). Fitting Multivariate Models to Community Data: a Comment on Distance-Based Redundancy Analysis. *Ecology* 82, 290–297. doi:10.1890/0012-9658(2001)082[0290:fmmtcd]2.0.co;2

Nguyen, N. P., Warnow, T., Pop, M., and White, B. (2016). A Perspective on 16s Rrna Operational Taxonomic Unit Clustering Using Sequence Similarity. *NPJ Biofilms Microbiomes* 2, 16004–16008. doi:10.1038/npjbiofilms.2016.4

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A Powerful and Adaptive Association Test for Rare Variants. *Genetics* 197, 1081–1095. doi:10.1534/genetics.114.165035

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential Abundance Analysis for Microbial Marker-Gene Surveys. *Nat. Methods* 10, 1200–1202. doi:10.1038/nmeth.2658

Rao, C. R. (1982). Diversity and Dissimilarity Coefficients: a Unified Approach. *Theor. Popul. Biol.* 21, 24–43. doi:10.1016/0040-5809(82)90004-1

Schenck, L. P., Surette, M. G., and Bowdish, D. M. E. (2016). Composition and Immunological Significance of the Upper Respiratory Tract Microbiota. *FEBS Lett.* 590, 3705–3720. doi:10.1002/1873-3468.12455

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Simpson, E. H. (1949). Measurement of Diversity. *Nature* 163, 688. doi:10.1038/163688a0

Sokal, R. R., and Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon* 11, 33–40. doi:10.2307/1217208

Song, C., Min, X., and Zhang, H. (2016). The Screening and Ranking Algorithm for Change-Points Detection in Multiple Samples. *Ann. Appl. Stat.* 10, 2102–2129. doi:10.1214/16-AOAS966

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* 449, 804–810. doi:10.1038/nature06244

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An Adaptive Association Test for Microbiome Data. *Genome Med.* 8, 56. doi:10.1186/s13073-016-0302-3

Yu, G., Lam, T. T.-Y., Zhu, H., and Guan, Y. (2018). Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol. Biol. Evol.* 35, 3041–3043. doi:10.1093/molbev/msy194

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in Microbiome-Profiling Studies with Mirkat, the Microbiome Regression-Based Kernel Association Test. *Am. J. Hum. Genet.* 96, 797–807. doi:10.1016/j.ajhg.2015.04.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership