STATISTICAL AND COMPUTATIONAL METHODS FOR MICROBIOME MULTI-OMICS DATA

EDITED BY: Himel Mallick, Vanni Bucci and Lingling An
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-091-9
DOI 10.3389/978-2-88966-091-9

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: researchtopics@frontiersin.org

STATISTICAL AND COMPUTATIONAL METHODS FOR MICROBIOME MULTI-OMICS DATA

Topic Editors:
Himel Mallick, Merck (United States), United States
Vanni Bucci, University of Massachusetts Dartmouth, United States
Lingling An, University of Arizona, United States

Image: paulista/Shutterstock.com

Citation: Mallick, H., Bucci, V., An, L., eds. (2020). Statistical and Computational Methods for Microbiome Multi-Omics Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-091-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 25 August 2020
doi: 10.3389/fgene.2020.00927

Editorial: Statistical and Computational Methods for Microbiome Multi-Omics Data

Himel Mallick[1]*, *Vanni Bucci*[2] and *Lingling An*[3,4,5]

[1] Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, United States, [2] Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, MA, United States, [3] Interdisciplinary Program in Statistics and Data Science, The University of Arizona, Tucson, AZ, United States, [4] Department of Epidemiology and Biostatistics, The University of Arizona, Tucson, AZ, United States, [5] Department of Biosystems Engineering, The University of Arizona, Tucson, AZ, United States

Keywords: microbiome, metagenomics, metabolomics, multi-omics, biostatistics, computational biology, statistical genomics, data science

Editorial on the Research Topic

Statistical and Computational Methods for Microbiome Multi-Omics Data

There has never been a more exciting time to do microbiome research thanks to the recent completion of several population-scale, longitudinal multi-omics studies including the NIH integrative human microbiome project (iHMP; iHMP Consortium, 2019) that have facilitated a multitude of new avenues of research for future investigations. These breakthroughs utilizing multiple 'omics technologies have paved the way toward investigating biological systems at an unprecedented level of detail, allowing a simultaneous assessment of community function, dynamics, and biochemical signatures across diverse disease states and environments. The field of microbiome multi-omics, however, has not yet reached the maturity attained in other established molecular epidemiology fields such as cancer biomarker discovery and genome-wide association studies (Mallick et al., 2017). As a result, it remains wide open to an in-depth exploration of new analytical methods in order to make the leap from bench to bedside.

This Research Topic is a timely endeavor toward this goal to expand our knowledge on systems biology approaches in understanding microbial communities. Due to the complexity of the associated data, the downstream analysis of microbiome multi-omics remains challenging. While most of the initial studies focused on analyzing single omics (e.g., taxonomic or functional profiles), there has been a shift in the field toward the concurrent investigation of the microbiome and host phenotypes (e.g., metabolomics and host transcriptomics). To this end, many of the articles in this Research Topic focus on new ways to analyze and integrate multi-table data using cutting-edge statistical and computational methods.

Sankaran and Holmes revisit an overwhelmingly large literature and algorithms already available on multi-table data analysis by reviewing both the algorithmic foundations and practical applications of a wide range of analysis approaches and re-evaluate these paradigms with respect to heterogeneity, dimensionality, and sparsity in a fully reproducible setup. In a similar vein, Bodein et al. propose a computational framework to integrate longitudinal microbiome data with other omics and clinical data generated on the same biological specimens based on smoothing splines and multivariate dimension reduction methods. Both these constitute a critical contribution to the field, given the growing commonality of multi-table datasets and the complexity of related study

OPEN ACCESS

Edited and reviewed by:
Simon Charles Heath, Center for Genomic Regulation (CRG), Spain

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

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 July 2020*
Accepted: *24 July 2020*
Published: *25 August 2020*

Citation:
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

designs, including dietary, pharmaceutical, clinical, and environmental covariates, often with samples from multiple time points or tissues.

Many important questions on microbiome multi-omics data integration remain unaddressed, especially those relating to extracting disease-relevant mechanistic networks that can provide insight into the complex web of host-microbiome interactions. Jiang et al. extensively review statistical aspects of relevant microbiome multi-omics network analysis methods by demystifying each class of methods with respect to their practical applicability and biological interpretability. Zhou and Gallins present a tutorial overview of commonly-used machine learning methods for microbiome host trait prediction, accompanied by validated R/Python implementations. The open-access source codes from these publications not only provide an important resource for algorithm developers but also ensure widespread usage and impact of these methods, facilitating future methodological research advances.

Moving beyond routine univariate analysis methods that ignore the correlations between features, Banerjee et al. take a multivariate approach to differential abundance analysis by jointly modeling all features in a set while maintaining the correct type I error and high power, which is not trivial for many existing per-feature methods (McMurdie and Holmes, 2014; Mandal et al., 2015; Jonsson et al., 2016, 2017; Thorsen et al., 2016; Mallick et al., 2017; Weiss et al., 2017; Hawinkel et al., 2019). Koh et al. introduce a distance-based kernel association test for family-based or longitudinal microbiome studies to associate microbial community composition with any type of host traits based on the generalized linear mixed model, vastly expanding the capability to incorporate non-Gaussian host traits as well as multiple kernels.

Quantitative methods of microbiome multi-omics are by no means limited to downstream analysis of targeted amplicon-based and metagenomic profiling. This Research Topic also contains papers addressing important questions in upstream data processing and quantitative microbiome profiling. For instance, Song et al. focus on the comparison of metagenomic samples using alignment-free methods with reads binning and conclude that alignment-free and alignment-based methods for metagenome comparison complement each other and should be used interactively to understand the dynamics of microbial communities. Yoon et al. estimate feature-feature correlations and partial correlations from robust measurements of microbial cell count, in particular, flow cytometry, and validate the results in a recent quantitative gut microbiome dataset ensuring both statistical rigor and biological relevance.

Several articles in the Research Topic go beyond integrating multiple omics datasets to establishing causation and molecular mechanism, with an emphasis on methods that aim to detect microbiome-mediated signals through causal mediation analysis. While existing methods in this space make strong parametric assumptions, which can be quite detrimental when the assumptions are violated, Carter et al. turn to nonparametric entropy models to detect significant mediation effects in the presence of high-dimensional exposures and mediators. Tang et al. utilize state-of-the-art microbiome compositional mediation analysis procedures to investigate the diet-microbiome-metabolome interaction in cross-sectional multi-omics samples from healthy subjects. Both these analyses estimate the total mediation effects of microbiome composition, as well as feature-specific mediation effects, providing additional mechanistic insights above and beyond a direct causal relationship.

Taken together, the papers in this Research Topic represent both an incredible amount of progress and an enormous potential for further advances in the near future. As a result, we have launched a second edition of the Research Topic where we will continue to add additional methods, research, and review articles over the next year or so.

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

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. doi: 10.1093/bib/bbx104

iHMP Consortium (2019). The integrative human microbiome project. *Nature* 569, 641–648. doi: 10.1038/s41586-019-1238-8

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* 17:78. doi: 10.1186/s12864-016-2386-y

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2017). Variability in metagenomic count data and its influence on the identification of differentially abundant genes. *J. Comput. Biol.* 24, 311–326. doi: 10.1089/cmb.2016.0180

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

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. doi: 10.3402/mehd.v26.27663

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4:62. doi: 10.1186/s40168-016-0208-8

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fgene.2019.00350

An Adaptive Multivariate Two-Sample Test With Application to Microbiome Differential Abundance Analysis

*Kalins Banerjee[1], Ni Zhao[2], Arun Srinivasan[3], Lingzhou Xue[3], Steven D. Hicks[4], Frank A. Middleton[5], Rongling Wu[1] and Xiang Zhan[1]**

[1] Department of Public Health Sciences, Pennsylvania State University, Hershey, PA, United States, [2] Department of Biostatistics, Johns Hopkins University, Baltimore, MD, United States, [3] Department of Statistics, Pennsylvania State University, University Park, PA, United States, [4] Department of Pediatrics, Pennsylvania State University, Hershey, PA, United States, [5] Department of Neuroscience, State University of New York Upstate Medical University, Syracuse, NY, United States*

Differential abundance analysis is a crucial task in many microbiome studies, where the central goal is to identify microbiome taxa associated with certain biological or clinical conditions. There are two different modes of microbiome differential abundance analysis: the individual-based univariate differential abundance analysis and the group-based multivariate differential abundance analysis. The univariate analysis identifies differentially abundant microbiome taxa subject to multiple correction under certain statistical error measurements such as false discovery rate, which is typically complicated by the high-dimensionality of taxa and complex correlation structure among taxa. The multivariate analysis evaluates the overall shift in the abundance of microbiome composition between two conditions, which provides useful preliminary differential information for the necessity of follow-up validation studies. In this paper, we present a novel **A**daptive multivariate two-sample test for **M**icrobiome **D**ifferential **A**nalysis (**AMDA**) to examine whether the composition of a taxa-set are different between two conditions. Our simulation studies and real data applications demonstrated that the AMDA test was often more powerful than several competing methods while preserving the correct type I error rate. A free implementation of our AMDA method in R software is available at https://github.com/xyz5074/AMDA.

Keywords: adaptive microbiome differential analysis (AMDA), maximum mean discrepancy (MMD), multivariate two-sample test, permutation, subset testing, taxa-set

OPEN ACCESS

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

Reviewed by:
Michael B. Sohn,
University of Rochester, United States
Hongmei Jiang,
Northwestern University, United States

****Correspondence:***
Xiang Zhan
xyz5074@psu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *08 January 2019*
Accepted: *01 April 2019*
Published: *24 April 2019*

Citation:
Banerjee K, Zhao N, Srinivasan A, Xue L, Hicks SD, Middleton FA, Wu R and Zhan X (2019) An Adaptive Multivariate Two-Sample Test With Application to Microbiome Differential Abundance Analysis. Front. Genet. 10:350. doi: 10.3389/fgene.2019.00350

1. INTRODUCTION

The human microbiome, referred as the aggregate of microorganisms that resides on or within any human tissues and biofluids, has recently gained substantial scientific interest due to its vital role in many human health and disease conditions, including but are not limited to obesity (Turnbaugh et al., 2009), type 2 diabetes (Qin et al., 2012), rheumatoid arthritis (Zhang et al., 2015), inflammatory bowel disease (Morgan et al., 2015), bacterial vaginosis (Mitchell et al., 2017), and colorectal cancer (Louis et al., 2014). High-throughput sequencing technologies have revolutionized microbiome research by allowing culture-free profiling of entire microbiome community. For the most part, 16S rRNA gene amplicon sequencing and metagenomics shotgun

sequencing are routinely used for quantitative characterization of microbiome composition (Wang and Jia, 2016). Although data produced by high-throughput sequencing has been proven extremely useful for quantification of microbiome composition, yet appropriate analysis of such microbiome composition data is still computationally and statistically challenging due to some technical aspects of the data, including high-dimensionality, count or compositional data structure, sparsity (zero-inflation), over-dispersion, among others.

In many microbiome studies, the investigators are often interested in studying how the abundance of microbiome is related with clinical characteristics of the samples, such as health/disease status, smoking status, or dietary habit (high-calorie or low-calorie). That is, many studies attempt to detect differentially abundant microbiome features (species/OTUs) between two predefined classes of samples, where a microbiome feature is considered differentially abundant, if its mean proportion is significantly different between two conditions. This type of analysis can improve understanding the pathology of the disease from a microbiome perspective and potentially lead to preventive or therapeutic strategies (Virgin and Todd, 2011). Microbiome differential abundance analysis (MDA) is a direct analogy to differential expression analysis for gene expression and RNA-seq data, however, the distinct nature of microbiome data renders classic differential expression analysis methods such as DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) inappropriate for microbiome data (McMurdie and Holmes, 2014; Weiss et al., 2017). Thus, new statistical methods for microbiome differential abundance analysis are desired.

Similar to individual gene-based and pathway-based differential expression analysis, there are two types of microbiome differential analyses: individual taxon-based univariate analysis and taxa set-based multivariate analysis. Along with the recent huge scientific interest in microbiome studies, many statistical methods for microbiome differential analysis have also been proposed (Sohn et al., 2015; Zhao et al., 2015; Zhang et al., 2016; Chen et al., 2017), with most of them focus on examining whether a single taxon is differentially abundant between two different conditions, followed by multiple testing correction methods adjusting for individual taxon p-values (e.g., the Benjamini-Hochberg/BH procedure, Benjamini and Hochberg, 1995). The control of False Discovery Rate (FDR) is necessary, as an excess of false discoveries may lead to costly follow-up validation studies on false positive taxa, which essentially are not differentially abundant. Despite their potential usefulness in identifying differentially abundant taxa, these individual analyses may suffer from the following inherent limitations. First, the type I error of an individual microbiome differential analysis may not be correct (Hawinkel et al., 2017). The BH procedure or its variant can control FDR when individual tests are either independent or under positive dependence assumptions (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), while negative correlation among taxa abundance is common in microbiome data, especially for compositional data. It is possible that these BH procedures (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) may fail to control FDR in presence of negative correlations (Hawinkel et al., 2017). Second, the high-dimensionality nature of microbiome data increases multiple correction burden of individual analyses, which reduces the power of detecting differentially abundant taxa. Third, as widely observed in literature, the performance of most individual microbiome differential analysis methods heavily rely on the normalization and/or transformation, leading to challenges in independent replication studies (McMurdie and Holmes, 2014; Sohn et al., 2015; Weiss et al., 2017).

An alternative approach to taxon-level microbiome differential analysis is to compare the microbiome composition at the level of taxa-set. Examples of such a taxa set can be either a group of OTUs belonging to the same upper-level taxonomic rank (e.g., phylum, class, order, family, or genus) or even all OTUs in the microbiome community. The multivariate-type microbiome differential analysis usually gains power by reducing the multiple testing correction burden and aggregating modest effects across multiple taxa. Moreover, the multivariate analysis is typically less sensitive to normalization/transformation compared to individual analysis as it has a much larger analysis unit. Motivated by this, many statistical methods for microbiome community-level analysis have been recently proposed (McArdle and Anderson, 2001; Zhao et al., 2015; Tang et al., 2016, 2017; Plantinga et al., 2017; Zhan et al., 2017a).

Despite of the potential power gain, a major critique of these existing multivariate microbiome analyses (e.g., differential analysis) is that the result of the test is global and is unable to identify specific taxon in the taxa-set that are differentially abundant. Besides the limitation in results' interpretation, it may also jeopardize the power of the test when the taxa-set contains many taxa that are not differentially abundant (Cao et al., 2017). To enhance both interpretation and power of existing multivariate analysis in the framework of MDA, we propose a two-stage Adaptive Microbiome Differential Analysis (AMDA) procedure, which first selects some putative taxa that are more likely to be differentially abundant between two conditions, and then examines the differential abundances of the selected taxa-set with a multivariate two-sample test using Maximum Mean Discrepancy (MMD) (Gretton et al., 2007, 2012). Since the test is applied to a subset of taxa that are more likely to be differentially abundant, permutations are used to establish statistical significance to avoid inflated type I error. Despite being a set-based multivariate test that does not target at identifying individual differentially abundant microbial taxa, the intermediate testing subset selection procedure in AMDA can provide useful information regarding the importance of individual taxon in the taxa-set. Simulation studies and real data applications demonstrate the potential usefulness of the new proposed AMDA method and show its superior performance over existing methods across a wide range of scenarios.

2. MATERIALS AND METHODS

2.1. Data and Normalization

Assume that we have measured the microbiome abundances of a community of p taxa from $n(= n_1 + n_2)$ samples collected from two groups with sizes of n_1 and n_2, respectively. Here, the

term community refers as a taxa-set, which typically consists of taxa from the same taxonomic rank such as genus, family, phylum, or bacteria kingdom. Let $\boldsymbol{X}^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})^T$ be the observed $n_k \times p$ OTU matrix for group $k(k = 1, 2)$, where $X_i^{(k)}(i = 1, \ldots, n_k; k = 1, 2)$ represents a $p \times 1$ microbiome composition vector (subject to appropriate normalization or transformation). Suppose that, $X_1^{(k)}, \ldots, X_{n_k}^{(k)}(k = 1, 2)$ are two independent samples, from p-dimensional multivariate distribution with mean parameters $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$, respectively. In many practical problems, the hypothesis of interest is to examine whether microbiome abundances are different under two different conditions, that is,

$$H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} \ \ vs. \ \ H_1 : \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)}. \tag{1}$$

For microbiome data, due to the varying amount of DNA yielding materials across different samples, the count of microbiome sequencing reads can vary greatly from sample to sample. The normalization of the raw sequencing read counts to relative abundances makes the microbial abundances comparable across samples. Therefore, it is a common practice to analyze high-dimensional microbiome compositional data with a unit sum (Li, 2015). As such, applying standard statistical methods developed for unconstrained data to analyze microbiome composition data is usually underpowered and sometimes can render inappropriate results (Cao et al., 2017; Weiss et al., 2017).

A popular approach to relax the compositional constraint of microbiome data is to perform the statistical analysis through log-ratio transformations (Aitchison, 1982). In particular, the centered log-ratio transformation has been widely used among various form of log-ratio transformations (Cao et al., 2017; Zhao et al., 2018). Specifically, the centered log-ratio transformation $Z_{ij}^{(k)}$ of microbiome relative abundance $X_{ij}^{(k)}$ is defined as

$$Z_{ij}^{(k)} = \log\left(\frac{X_{ij}^{(k)}}{(\Pi_{j=1}^{p} X_{ij}^{(k)})^{1/p}}\right), \quad i = 1, \ldots, n_k, j = 1, \ldots, p, \\ k = 1, 2. \tag{2}$$

To avoid a zero relative abundance in Equation (2), as a common practice, a zero count is usually replaced by a pseudo count of 0.5 before the relative abundance normalization and centered log-ratio transformation (Li, 2015; Cao et al., 2017). For community-based multivariate differential abundance analysis, it has been shown that testing equality of two compositional vectors is equivalent to testing $H_0' : \boldsymbol{\mu}_Z^{(1)} = \boldsymbol{\mu}_Z^{(2)}$ (Cao et al., 2017), where $\boldsymbol{\mu}_Z^{(k)}$ is the mean of centered log-ratio transformed compositional vector $Z_i^{(k)}$, $i = 1, \ldots, n_k$ and $k = 1, 2$. We will develop our AMDA method based on these centered log-ratio transformed relative abundances in the rest of this paper.

2.2. A Multivariate Two-Sample Test Using Maximum Mean Discrepancy

Two-sample testing on the equality of two high-dimensional means has been well studied in the statistical literature (Bai and Saranadasa, 1996; Chen et al., 2010; Cai et al., 2014). These methods are typically not applicable to MDA analysis due to the following two reasons. First, existing methods usually assume normal data, which is not the case for microbiome compositional data. It has been observed that classic statistical methods developed for multivariate Gaussian data may fail for microbiome compositional data (Li, 2015; Cao et al., 2017; Zhao et al., 2018). Second, most existing methods require estimating the covariance matrix. Given the small or modest sample size in a typical microbiome study, the relatively large estimation error of covariance matrix probably deteriorates the performance of two-sample test, as observed in microbiome association tests (Zhan et al., 2017b, 2018).

An alternative approach to test hypothesis (Equation 1)is to use a non-parametric test that does not need to estimate the covariance matrix. One such test is the kernel-based maximum mean discrepancy (MMD) test (Gretton et al., 2007, 2012), originally proposed to examine whether the underlying distribution of two samples are identical. An MMD test first maps the two distributions into a reproducing kernel Hilbert space (RKHS) and then the maximum mean discrepancy metric between the two distributions is defined as the distance of their corresponding images in the RKHS. A good property about MMD is that, MMD is zero if and only if two distributions are identical when the RKHS is sufficiently rich (contain a large enough class of functions). Since the test can be used to examine equality of two multivariate distributions, it suffices for testing (Equation 1), that is, to examine the equality of the mean parameters of two underlying distributions.

In particular, the MMD statistic between two independent samples $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ is defined as

$$\mathrm{MMD}^2 = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} k(X_i^{(1)}, X_j^{(1)}) + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} k(X_i^{(2)}, X_j^{(2)}) \\ - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} k(X_i^{(1)}, X_j^{(2)}), \tag{3}$$

where $k(\cdot, \cdot)$ is a characteristic kernel (Gretton et al., 2007, 2012), which spans a RKHS which is sufficiently large that MMD is zero if and only if two samples are from the same underlying distribution. Examples of characteristic kernel include the Gaussian kernel and the Laplace kernel. Under the null hypothesis of identical distribution, the population-level MMD^2 statistic is zero, and thus, a larger MMD^2 statistic indicates a larger discrepancy between the two distributions. Asymptotically, MMD^2 follows a mixture of χ_1^2 distribution (Gretton et al., 2007, 2012). As observed in literature, the asymptotic mixture of χ_1^2 distribution is typically not accurate for a statistic calculated from a small sample size, as frequently encountered in microbiome studies (Chen et al., 2016; Zhan et al., 2017b, 2018). A more accurate approach to establish significance is using resamplings (e.g., permuting the group label of each observation) (Wu et al., 2016).

2.3. An Adaptive Two-Sample Test for Microbiome Differential Abundance Analysis

A limitation of the aforementioned MMD test is that it equally utilizes information in all dimensions. When the signal is sparse, the MMD test typically has a low power due to the high degrees of freedom paid for many noise variables. The same phenomenon has been widely observed in the field of set-based genetic association studies (Cai et al., 2012; Pan et al., 2014, 2015; Zhan et al., 2015) and community-based microbiome association studies (Wu et al., 2016; Koh et al., 2017). There are in general two types of two-sample test of high-dimensional means. One is based on the sum of squares of mean differences of each dimension [e.g., MiRKAT proposed in Zhao et al., 2015], and the other is based on the largest componentwise mean difference (e.g., the max-type test proposed in Cao et al. (2017)). For microbiome differential abundance analysis, the max-type test tends to be more powerful when only a few taxa are truly differentially abundant. On the other hand, the MiRKAT-type test can be more powerful than the max-type test under the scenario of dense signals. In practice, the true underlying biological scenario is never known and thus adaptive methods for microbiome differential abundance analysis are desired.

A common adaptive approach in a multivariate association test or two-sample test is to assign different weights to variables so that important variables are up-weighted and non-informative variables are down-weighted (Cai et al., 2012; Pan et al., 2014, 2015; Wu et al., 2016; Koh et al., 2017). Yet it is often difficult to determine the optimal weights. Some authors propose another loop of permutations to combine multiple sets of weights, which may be computationally challenging since most adaptive tests already need permutations to establish significance (Pan et al., 2014, 2015). In this paper, we propose a different adaptive method, which tests the hypothesis in a selected subset of microbiome features. In other words, instead of applying the MMD test to all p taxa $\boldsymbol{X} = (X_1, \ldots, X_p)$, we apply the test on a putative testing subset $\boldsymbol{X}_S$, where $S \subset \{1, \ldots, p\}$. Our method can also be viewed as a weighted approach in the sense that a zero weight is assigned to a feature that is not selected in the testing subset, and an equal weight is assigned to each feature in the testing subset. We defer details of selecting such a testing subset to the next section and present our adaptive microbiome differential analysis (AMDA) procedure in **Algorithm 1**:

Algorithm 1: An adaptive two-sample test for microbiome differential abundance analysis

Input: A $n \times p$ microbiome composition matrix $\boldsymbol{X} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)})^T$ and a $n \times 1$ group label vector $y = (1, \ldots, 1, 2, \ldots, 2)$ associated with the microbiome compositions.
Output: A p-value for $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ *vs.* $H_1 : \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)}$.
Procedure:

1. Apply the centered log-ratio transformation Equation (2) to the microbiome composition matrix. Without loss of generality, we still use $\boldsymbol{X}$ to denote the centered log-ratio transformed data.
2. Use the testing subset selection procedure described in section 2.4 to select a testing subset $\boldsymbol{X}_{\mathcal{S}}$ from $\boldsymbol{X}$, and then calculate the MMD statistic using $\boldsymbol{X}_{\mathcal{S}}$ and y. Denote this statistic as MMD^2_{obs}.
3. For $b = 1, \ldots, B$, permute the group label of observations to obtain $\tilde{y}$ and use $\tilde{y}$ to repeat Step 2 with $\boldsymbol{X}$ and $\tilde{y}$. Calculate the corresponding statistics as MMD^2_b for $b = 1, \ldots, B$.
4. Calculate the final p-value as $pv = \frac{1}{B}\sum_{b=1}^{B} I[MMD^2_b \geq MMD^2_{obs}]$, where $I[\cdot]$ is the indicator function.

2.4. A New Permutation-Based Testing Subset Selection Procedure

There is a vast statistical literature on high-dimensional variable selection. Some famous examples include the lasso (Tibshirani, 1996) and the knockoff filter (Barber and Candès, 2015; Candes et al., 2018). The lasso has proven to be a versatile tool with nice asymptotic estimation and prediction properties, yet its performance under small sample size is not guaranteed. On the other hand, knockoff is able to select variables under FDR control with finite samples. But it tends to select a smaller set of variables with less false positives to achieve FDR control (see **Table S1** in the online supplemental material). As a consequence, many signals are not selected by knockoff, typically leading to a less powerful test. Recall that, our ultimate goal is to construct a differential test with relatively high power. For this reason, we prefer a procedure that can select a testing subset that contains as many signals as possible. To achieve this goal, we propose the following permutation-based testing subset selection procedure.

We first randomly permute the row indices of matrix $\boldsymbol{X}$ (defined in **Algorithm 1**) and obtain a permuted microbiome composition matrix $\tilde{\boldsymbol{X}}$. By the nature of its construction, $\tilde{\boldsymbol{X}}$ is not related to outcome y. Next, a one-dimensional two-sample test (e.g., the Kolmogorov-Smirnov test) is applied to each dimension of $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, and we denote the corresponding p-values as $p_1, \ldots, p_p$ and $\tilde{p}_1, \ldots, \tilde{p}_p$, respectively. Because the dimension p is typically much larger than sample size in microbiome studies, we calculate the marginal p-values rather than joint p-values for testing subset selection. For a truly differentially expressed variable X_j, as $\tilde{X}_j$ is not constructed to be outcome-related, it is expected that $p_j < \tilde{p}_j$. Hence, we select the testing subset as $\mathcal{S} = \{j : p_j < \tilde{p}_j\}$ and conduct our MMD test based on the sub-design matrix $\boldsymbol{X}_{\mathcal{S}}$. Finally, as we are testing $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ using microbiome features that are more likely to be differentially expressed, to avoid inflated type I error, resampling methods are required to establish the significance (see details in **Algorithm 1**).

It should be noted that the aforementioned permutation-based procedure is one way to achieve testing subset selection but not the only way, and it is possible to select testing subset $\boldsymbol{X}_{\mathcal{S}}$ using other methods such as lasso and knockoff. We conduct comprehensive simulation studies to compare the power of adaptive two-sample test using different testing subset selection procedures and report the results in the online **Supplementary Material**. As can be observed there, adaptive test based on our permutation-based procedure is more powerful

than both lasso-based and knockoff-based tests, as both lasso and knockoff tend to miss more true signals for the sake of achieving sparsity (lasso) or FDR control (knockoff).

3. RESULTS

3.1. Simulation Settings

A comprehensive simulation study has been conducted to compare the performance of AMDA to a wide range of existing microbiome association tests in the framework of microbiome differential abundance analysis. The five other tests evaluated in this simulation include the MiRKAT (Zhao et al., 2015), the original MMD test without testing subset selection (Gretton et al., 2007, 2012), the Quasi-Conditional Association Test/QCAT (Tang et al., 2017), the maximum-type (MAX) test based on the largest sample mean difference (Cao et al., 2017) and the optimal microbiome-based association test/OMiAT (Koh et al., 2017). AMDA, MiRKAT, MMD, QCAT, and MAX are a single test, while OMiAT takes advantage of two series of tests. One is the MiSPU tests (Wu et al., 2016) with different weighting schemes on each individual taxon in the taxa-set. The other is the MiRKAT tests with different kernel functions. The spirit of OMiAT can be easily implemented in AMDA, MiRKAT, and MMD by evaluating multiple kernels and taking the optimal kernel test with minimum p-value. We do not incorporate this strategy, for ease of presenting, and only evaluate the Gaussian kernel-based test for AMDA, MiRKAT, and MMD in this simulation. Correspondingly, we evaluate the OMiAT as the optimal of a series of MiSPU tests (without MiRKAT tests of different kernels) for fair comparison. With a slight abuse of notation, we still term this test as OMiAT, though it does not contain the MiRKAT component compared to the original one (Koh et al., 2017). Moreover, QCAT and MAX tests with asymptotic p-values are found to have inflated type I errors (data not shown). For this reason, we use permutations to calculate the MAX test p-value and the resampling option in the QCAT software (Tang et al., 2017) to calculate QCAT p-value. Finally, the permutation-based procedure is used to select testing subset in the intermediate stage of AMDA in this simulation. The performance of AMDA test based on other subset selection methods such as lasso and knockoff were evaluated in additional simulation studies presented in the online **Supplementary Material**.

We closely followed the simulation design of the MAX test (Cao et al., 2017) to generate microbiome relative abundances data using the logistic normal distribution (Atchison and Shen, 1980). We first simulated $\boldsymbol{W}_i^{(k)} \sim N_p(\boldsymbol{\mu}^{(k)}, \Sigma)$ for $i = 1, 2, \ldots, n$, $k = 1, 2$ and then calculated the microbiome relative abundances as $X_{ij}^{(k)} = \exp[W_{ij}^{(k)}] / \sum_{j=1}^{p} \exp[W_{ij}^{(k)}]$ and its centered log-ratio transformation $Z_{ij}^{(k)}$ according to Equation (2). Following the simulation design of MAX (Cao et al., 2017), the components of $\boldsymbol{\mu}^{(1)}$ were drawn from a uniform distribution Unif(0,10) and we considered the banded covariance structure $\Sigma = \boldsymbol{D}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}$, where $\boldsymbol{D}$ is a diagonal matrix with entries randomly drawn from Unif(1,3) and $\boldsymbol{A}$ has nonzero entries $a_{jj} = 1$, $a_{j,j-1} = a_{j-1,j} = -0.5$. Under the null model, we set $\boldsymbol{\mu}^{(2)} = \boldsymbol{\mu}^{(1)}$. Under the alternative model, we randomly picked a subset $\mathcal{S} \subset \{1, 2, \ldots, p\}$ such that $\mu_j^{(2)} = \mu_j^{(1)} + e_j$, where $e_j \sim Unif(-0.5, 0.5)$ for all $j \in \mathcal{S}$. For the size of signal set $\mathcal{S}$ (number of taxa that are truly differentially abundant), we considered low, medium and high signal density levels: $p^* = |\mathcal{S}| = 10\%p$, $30\%p$ and $50\%p$ with the indices randomly chosen from $\{1, 2, \ldots, p\}$. Throughout this simulation, we varied $n = 50, 100, 200$ with $n_1 = n_2 = n/2$ to investigate the test's performance under different sample sizes, and considered $p = 50, 100, 200, 500$ representing taxa-sets under different taxonomic ranks.

After the data were simulated, we applied AMDA, MAX, OMiAT, MMD, MiRKAT, and QCAT to examine the two-sample differences. The first three tests AMDA, MAX, OMiAT are adaptive in the sense that they either use a testing subset of the taxa (AMDA and MAX) or assign a different weight for each taxon in the set (OMiAT) to conduct the multivariate two-sample test. The Gaussian kernel ($k(x, y) = \exp\{-||x - y||^2/\rho\}$, where x and y are two microbiome compositional vectors) was used in AMDA, MMD, and MiRKAT with the shape parameter ρ selected as the median of sample pairwise Euclidean distance $||x - y||^2$. The type I error was evaluated using 5,000 replicates generated under the null model and the power of test was assessed with 1,000 replicates under the alternative model. Without loss of generality, we set the nominal significance level $\alpha = 0.05$ throughout this simulation.

3.2. Simulation Results

The type I error of different tests are reported in **Table 1**, where one can see that all tests have the correct type I error across all (n, p)-configurations. The power of different tests are reported **Figure 1** ($p = 50$ and 100) and **Figure 2** ($p = 200$ and 500). Since the effect size was arbitrarily chosen to avoid power saturation, we care about the relative power among different methods rather than their absolute magnitudes. As can be seen from both figures, adaptive tests (AMDA, MAX, and OMiAT) are consistently more powerful than the non-adaptive ones (MMD, MiRKAT, and QCAT). This is because the scenarios considered in our simulation studies are relatively sparse ($p^*/p \leq 50\%$), and the adaptive tests can largely boost the power by treating variables (signals and noises) differently.

Among three non-adaptive tests, MMD and MiRKAT have similar power under each scenario. On the other hand, QCAT has the highest power when the dimension of taxa-set is relatively low (**Figure 1**) especially when the sample size is relatively large ($n = 200$). When the dimension of taxa-set increases, QCAT can quickly lose power and become less powerful than both MMD and MiRKAT (**Figure 2**).

Among the three more powerful adaptive tests, MAX seems to be slightly more powerful than AMDA and OMiAT when the density of signal is sparse ($p^*/p = 10\%$) and dimension is relatively low ($p = 50$,100, and 200) as indicated in **Figure 1** and the top row of **Figure 2**. Compared to AMDA, MAX only utilizes the strongest signal, which could be beneficial when the signals are extremely sparse. When $p = 500$, there are $p^* = 50$ even under the sparse scenario and AMDA can be more powerful than MAX by including more signals in the testing subset (bottom row of **Figure 2**). On the other hand, when the signal level is moderate ($p^*/p = 30\%$)

TABLE 1 | Empirical type I errors of different tests for microbiome differential abundance analysis under nominal significance level $\alpha = 0.05$.

p	n	AMDA	MAX	OMiAT	MMD	MiRKAT	QCAT
	50	0.0478	0.0478	0.0506	0.0516	0.0508	0.0436
50	100	0.0464	0.0458	0.0492	0.0536	0.0540	0.0488
	200	0.0504	0.0542	0.0530	0.0534	0.0548	0.0480
	50	0.0486	0.0478	0.0490	0.0434	0.0424	0.0532
100	100	0.0464	0.0494	0.0492	0.0544	0.0542	0.0478
	200	0.0524	0.0558	0.0514	0.0440	0.0424	0.0470
	50	0.0454	0.0498	0.0492	0.0438	0.0400	0.0490
200	100	0.0514	0.0476	0.0464	0.0530	0.0516	0.0538
	200	0.0464	0.0510	0.0506	0.0542	0.0530	0.0476
	50	0.0480	0.0464	0.0504	0.0556	0.0442	0.0474
500	100	0.0540	0.0544	0.0566	0.0570	0.0498	0.0468
	200	0.0556	0.0576	0.0456	0.0490	0.0442	0.0336

Results are averaged over 5,000 replicates.

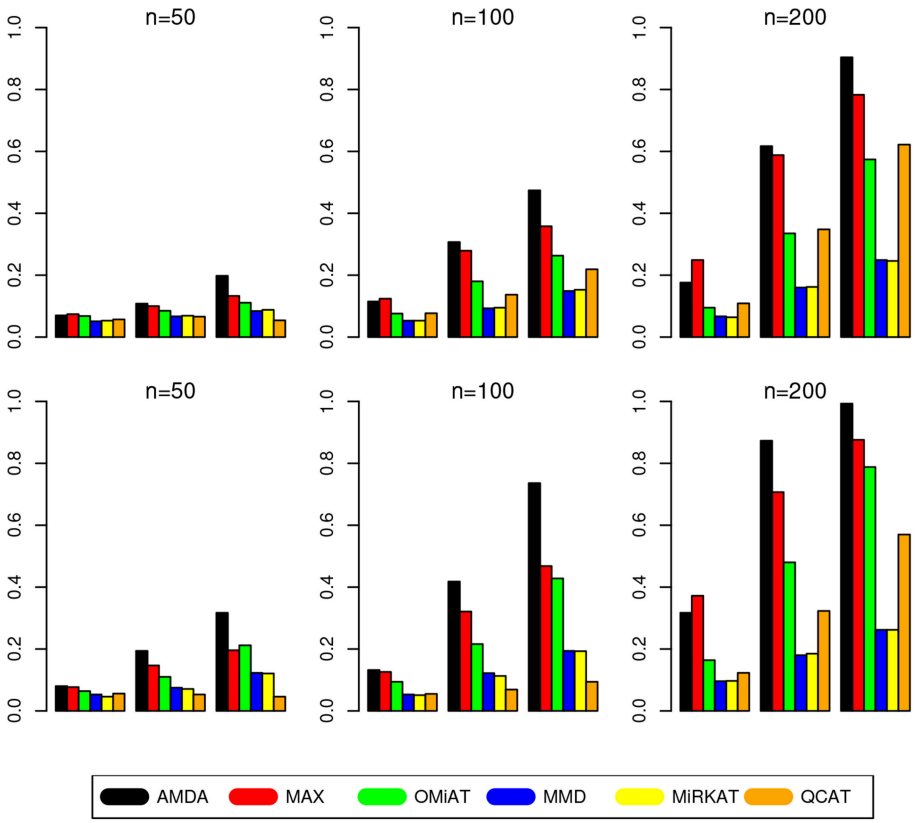


FIGURE 1 | Empirical power of different tests under $p = 50$ (first row) and $p = 100$ (second row). The Y-axis represents the power and the X-axis represents the sparsity level at 10, 30, and 50%.

or relatively dense ($p^*/p = 50\%$), AMDA is much more powerful than MAX under most scenarios in both **Figures 1**, **2**. Finally, as seen from both figures, AMDA is always more powerful than OMiAT across all scenarios. AMDA and OMiAT treat variables in different ways. AMDA selects some variables and excludes the rest for further subset testing, while OMiAT assigns different weights for different variables when calculating the multivariate score test statistic. Despite that a small non-zero weight may be assigned to a noise variable in OMiAT, due to the relatively sparse signal density ($p^*/p \leq 50\%$, which means there are much more noises than signals), the accumulated adverse effects of noise variables can still deteriorate the performance of OMiAT. As a comparison, a zero weight is assigned to a noise variable (by excluding it from the

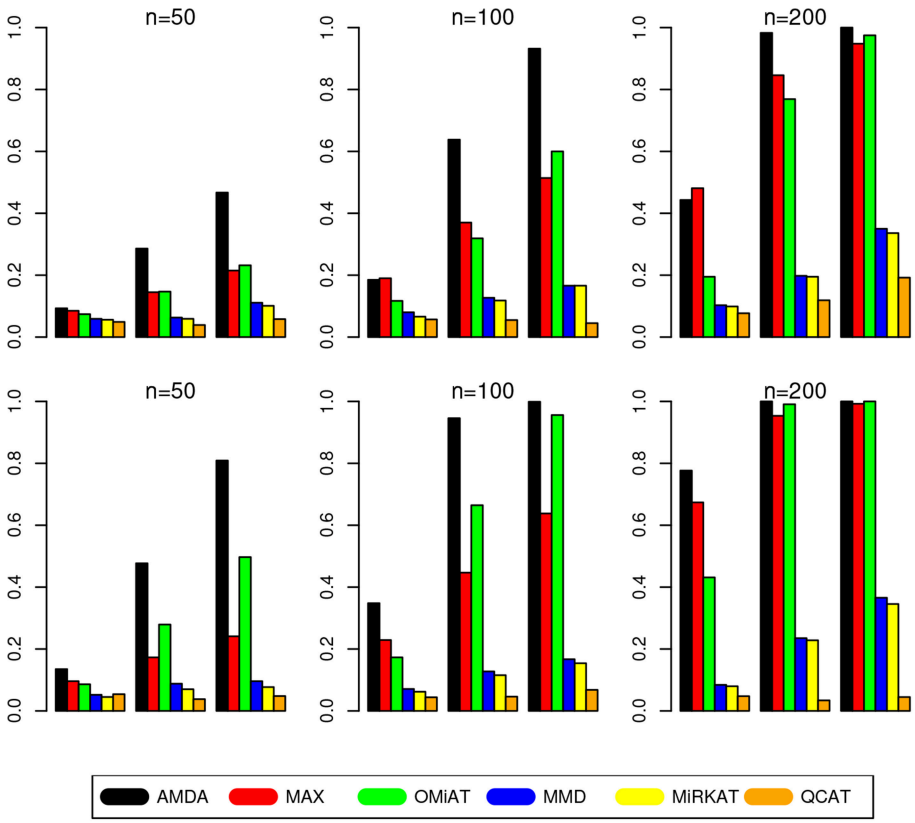


FIGURE 2 | Empirical power of different tests under $p = 200$ (first row) and $p = 500$ (second row). The Y-axis represents the power and the X-axis represents the sparsity level at 10, 30, and 50%.

testing subset) in AMDA, which explains power gain in AMDA over OiMAT.

To conclude, like five other methods, the proposed AMDA method is able to preserve the nominal type I error in microbiome differential abundance analysis. Power-wise speaking, there is no uniformly most powerful test in our simulations. However, the proposed AMDA method is always the most powerful one among all six tests being evaluated in this simulation under most scenarios, and the power advantage of AMDA over the other five methods can be huge (**Figures 1**, **2**). Under only a few particular scenarios with extremely sparse signal ($p^*/p = 10\%$) under relative low dimensions ($p = 50$,100, and 200), MAX can be slightly more powerful than AMDA.

3.3. Application to Oral Microbiome Data Collected From Children With Autism Spectrum Disorder

We applied the proposed AMDA method to a study investigating how the oral microbiome differs across children with autistic behaviors (Hicks et al., 2018). The study enrolled 346 children (between 2 and 6 years old), which were divided into three groups according to the severity of disorder/developmental status: autism spectrum disorder (ASD, $n = 180$), non-autistic developmental delay (DD, $n = 60$), and typically developing (TD, $n = 106$). The ASD group was defined using criteria specified in the Diagnostic and Statistical Manual of Mental Disorders (DSM–5) by the American Psychiatric Association. The DD group included children who did not meet DSM-5 criteria for ASD but had developmental delay symptoms (e.g., expressive speech delay and intellectual disability). TD children included children with negative ASD screening and met typical developmental milestones on standardized physician assessment. The oral microbiome composition of these children was quantified with next generation sequencing. The data along with details of data processing are available in the previous publication (Hicks et al., 2018).

Taxonomic reads were further filtered to include only the taxa with counts of more than 10, in more than 20% samples, which ended up with a oral microbiome community of 753 taxa. Sequence alignment with the k-SLAM (Ainsworth et al., 2017) method was used for comprehensive taxonomic classification, and these 753 taxa were classified into 457 species, 266 genera, 142 families, 73 orders, 33 classes, and 16 phyla (each rank had a Unclassified group for taxonomic sequence not identified at that rank). Because the proposed AMDA method is an adaptive multivariate two-sample test, we focused our analysis on higher taxonomic ranks (family, order, class, phylum, and the community of all 753 taxa), as many lower taxonomic ranks contain only a single taxon (e.g., 410 of the 457 species are

a singleton). Similarly, for the taxonomic ranks (family, order, class, and phylum) being considered, we further limited our analysis to a particular taxa-set that contains more than two taxa. As a result, 52 families, 34 orders, 18 classes, and 10 phyla were tested in our data analysis. We applied AMDA, MAX, OMiAT, MMD, MiRKAT, and QCAT to this data to examine the oral microbiome differences among three different children developmental profile groups (particularly, ASD vs. DD and ASD vs. TD) at different taxonomic ranks. As 52 families/34 orders/18 classes/10 phyla were tested, we adjusted for multiple testing using the Bonferroni correction to control the family-wise error rate at $\alpha = 0.05$. Correspondingly, $B = 10,000$ permutations/resamplings were used in AMDA, MAX, OMiAT, MMD, and QCAT to increase the precision of the test p-values, while the MiRKAT calculates the p-value analytically.

We first applied these tests to examine whether there is an overall shift in oral microbiome composition between different developmental groups by testing the differential abundances of all 753 taxa as a whole community. For the comparison of ASD vs. DD, the test p-values of AMDA, MAX, OMiAT, MMD, MiRKAT, and QCAT are 0.0113, 0.1409, 0.5244, 0.1321, 0.1377, and 0.9802, respectively. AMDA is the only method that is able to detect a significant (p-value < 0.05) difference of microbiome community profiles between ASD and DD. For the comparison of ASD vs. TD, the test p-values of AMDA, MAX, OMiAT, MMD, MiRKAT, and QCAT are 0.0021, 0.0017, 0.0323, 0.3039, 0.3099, and 0.1782, respectively. All three adaptive methods (AMDA, MAX, and OMiAT) are able to detect a significant difference between ASD and TD. In the original study (Hicks et al., 2018), the Mann-Whitney U-test based individual differential analysis was applied to each taxon and only three/six taxa were differentially abundant between ASD vs. DD/ASD vs. TD under FDR = 0.05 [see **Table 2** of Hicks et al. (2018)]. According to the previous simulation results, when the number of signals is relatively small ($p^* = 3$ or 6 as suggested in the original analysis) compared to the number of variables ($p = 753$), the non-adaptive tests have a low power. This explains that MMD/MiRKAT/QCAT methods are not able to detect a significant difference of microbiome profiles between two conditions in this data. Finally, the AMDA/MAX/OMiAT p-value of comparison ASD vs. TD is much smaller than that of comparison ASD vs. DD, indicating a more significant overall oral microbiome composition difference between ASD vs. TD than the between ASD vs. DD, which is consistent with the severity of disorder.

Next, we shift our analysis unit to lower ranks than the community-level to comprehensively assess taxa-set (with multiple taxa) at each taxonomic rank that are differentially abundant among different developmental status groups. The testing results are summarized here in **Table 2** . Based on this table, one can observe that the proposed AMDA always declares more significant differences than the other two tests except for one scenario (class-level differential analysis between ASD and TD). The absolute difference among three methods presented in **Table 2** may be small due to the conservativeness of the Bonferroni correction. To observe the relative trends of different tests, the p-values of these tests at family-level are presented in **Figure 3** (p-values at other taxonomic ranks have the similar pattern and hence are not reported). The AMDA p-values tend to be the smallest among p-values of all six tests. Therefore, our method has a clear advantage over the other methods in terms of detecting more significant differences in this oral microbiome data differential abundance analysis.

4. DISCUSSION

With the ever-increasing availability of microbiome and metagenomics data generated by next generation sequencing technology, the need to develop and implement efficient statistical analysis for the data is important to ensure both statistical rigor and biological relevance. In this paper, we consider the problem of differential abundance analysis for microbiome data, which leads to a better understanding of the behavior of microbiome communities. Most existing methods tackle this problem using individual taxon-based approach followed by multiple testing adjustment. However, as taxa living in the same community do not grow independently, the complicated interactions among taxa result in complicated correlation structures among taxa relative abundances, which may violate the correlation assumptions (among individual tests) of existing multiple correction methods (Hawinkel et al., 2017). On the other hand, the newly proposed AMDA examines the differential abundance of a taxa-set typically containing taxa from the same genus/family/order/class/phylum, which provides an invaluable compliment to the individual taxon-based differential abundance analysis. Given evidence of an association of a taxa-set with the outcome and assuming that at least one outcome-associated taxon within the set exist, applying AMDA to a high taxonomic rank can provide a useful preliminary screening of the whole microbiome (all species in the community) and facilitate more targeted downstream laboratory-based microbiome fine-mapping and functional studies (Wang and Jia, 2016).

The AMDA method has two main advantages compared to a traditional individual taxon-based approach. First, it can provide new biological and biomedical insights. The joint modeling of all taxa in the set is able to capture conditional effects of taxa that are missed in the traditional individual taxon-based approach, and thus new insights can be gained by shifting the analysis unit to a higher taxonomic rank. Second, it is statistically powerful by aggregating marginal signals of individual taxon and reducing the multiple testing burden. By adaptively choosing the subset being tested, our AMDA further boosts the statistical testing power compared to existing taxa set-based differential abundance analyses (e.g., MiRKAT). Moreover, the adaptive strategy used in AMDA could be easily extended to other hypothesis testing framework (e.g., association testing) beyond the two-sample problem considered in this paper. We conducted comprehensive numerical simulation studies to show the superior performance of AMDA over existing approaches in terms of maintaining the correct type I error while having a higher power to detect a true difference. The potential usefulness of AMDA was further

TABLE 2 | Number of significant differential abundant taxa-set at each taxonomic rank detected by different methods under family-wise error rate of 0.05.

Comparison	Rank	AMDA	MAX	OMiAT	MMD	MiRKAT	QCAT
ASD vs. DD	Phylum (10)	3	1	0	0	0	1
	Class (18)	3	1	1	2	2	1
	Order (34)	2	0	0	1	1	2
	Family (52)	1	0	0	0	0	1
ASD vs. TD	Phylum (10)	2	2	2	0	0	1
	Class (18)	4	3	3	2	2	5
	Order (34)	3	2	1	1	1	2
	Family (52)	2	2	1	2	2	2

Number in parentheses denotes the total number of tests conducted at that rank.

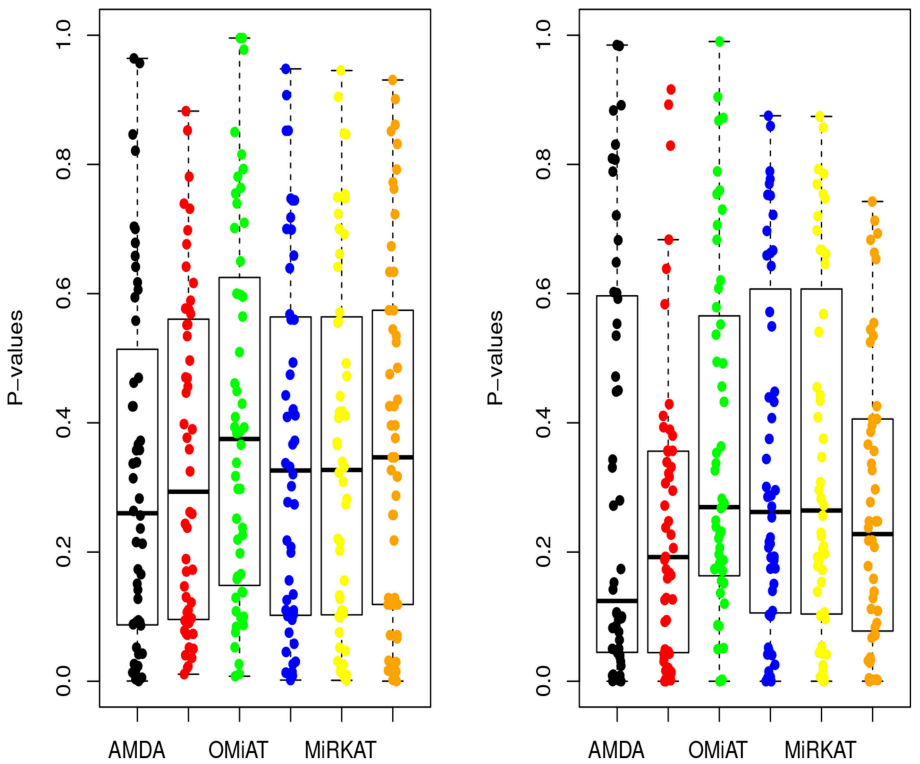


FIGURE 3 | *P*-values of AMDA, MAX, OMiAT, MMD, MiRKAT, and QCAT for family-level differential abundance analysis. The left panel corresponds to the comparison between ASD and DD, and right panel corresponds to the comparison between ASD and TD.

demonstrated via its application to an oral microbiome data, where AMDA tends to detect more significant differences than its competitors.

For illustration of our method, we applied the Gaussian kernel-based MMD test, which has been shown to be a consistent two-sample test (Gretton et al., 2007, 2012). The numerical performance of AMDA using other kernels including Unifrac and Bray-Curtis (Zhao et al., 2015) is similar to the one based on the Gaussian kernel (data not shown). As the field matures, more complex (such as family-based and longitudinal) study designs have become increasingly popular in the scientific community to study the association between microbiome and various clinical and biological covariates. This is partially because these advanced designs can be more efficient to control potential confounders compared to the population-based studies with unrelated individuals. The current adaptive multivariate microbiome differential abundance analysis is developed for independent samples. It is of further interest to extend it to accommodate correlated microbiome samples collected from a study using such a complex design. The current permutation-based testing subset selection procedure has been shown to have better numerical performance in terms of selecting more signals into testing subset than existing methods across a wide range of scenarios. Yet, any theoretical

guarantees of this permutation-based selection procedure is largely unknown. It is also of interest to further incorporate the phylogenetic tree information into AMDA to facilitate a comprehensive microbiome differential abundance analysis besides applying AMDA to one taxonnomic rank of the tree each time. We believe these issues are of importance and warrant further investigation.

ETHICS STATEMENT

This study involves only secondary analyses, where all the utilized data sets are published in a previous study.

AUTHOR CONTRIBUTIONS

KB and NZ analyzed the data, drafted the paper, prepared figures and tables, AS and LX conducted the testing subset simulations, SH and FM provided and helped analyze the oral microbiome data. RW contributed substantial expertise to improve the paper and revised the paper. XZ conceived and designed the experiments, analyzed the data, wrote the paper, and software. All authors read and approved the final manuscript.

FUNDING

This work was supported by Quadrant Biosciences Inc. (Research agreement with SH), the National Institutes of Health grants R41 MH111347 (FM), P50 DA039838 (LX) and National Science Foundation grant DMS-1811552 (LX).

ACKNOWLEDGMENTS

The authors would like to thank the Associate Editor and two reviewers for their insightful comments that improved the paper. Funding was provided by Quadrant Biosciences Inc. (Research agreement with SH) and NIH STAR (R41 MH111347).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00350/full#supplementary-material

REFERENCES

Ainsworth, D., Sternberg, M. J., Raczy, C., and Butcher, S. A. (2017). k-slam: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res.* 45, 1649–1656. doi: 10.1093/nar/gkw1248

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* 44, 139–177.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Atchison, J., and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* 67, 261–272.

Bai, Z., and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Stat. Sin.* 6, 311–329.

Barber, R. F., and Candès, E. J., (2015). Controlling the false discovery rate via knockoffs. *Ann. Stat.* 43, 2055–2085. doi: 10.1214/15-AOS1337

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser.* 57, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1214/aos/1013699998

Cai, T., Lin, X., and Carroll, R. J. (2012). Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics* 13, 776–790. doi: 10.1093/biostatistics/kxs015

Cai, T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc.* 76, 349–372. doi: 10.1111/rssb.12034

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: model–X knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc.* 80, 551–577. doi: 10.1111/rssb.12265

Cao, Y., Lin, W., and Li, H. (2017). Two-sample tests of high-dimensional means for compositional data. *Biometrika* 105, 115–132. doi: 10.1093/biomet/asx060

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

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., et al. (2017). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* 34, 643–651. doi: 10.1093/bioinformatics/btx650

Chen, S. X., and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.* 38, 808–835. doi: 10.1214/09-AOS716

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007). "A kernel method for the two-sample problem," in *NIPS* 513–520. Cambridge, MA: MIT Press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* 13, 723–773.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2017). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* 20, 210–221. doi: 10.1093/bib/bbx104

Hicks, S. D., Uhlig, R., Afshari, P., Williams, J., Chroneos, M., Tierney-Aves, C., et al. (2018). Oral microbiome activity in children with autism spectrum disorder. *Aut. Res.* 11, 1286–1299. doi: 10.1002/aur.1972

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. doi: 10.1186/s40168-017-0262-x

Li, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Ann. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Louis, P., Hold, G. L., and Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* 12:661. doi: 10.1038/nrmicro3344

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

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comp. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., et al. (2017). Vaginal microbiota and genitourinary menopausal symptoms: a cross-sectional analysis. *Menopause* 24, 1160–1166. doi: 10.1097/GME.0000000000000904

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Gen. Biol.* 16:67. doi: 10.1186/s13059-015-0637-x

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* 197, 1081-95. doi: 10.1534/genetics.114.165035

Pan, W., Kwak, I.-Y., and Wei, P. (2015). A powerful pathway-based adaptive test for genetic association with common or rare variants. *Am. J. Hum. Genet.* 97, 86–98. doi: 10.1016/j.ajhg.2015.05.018

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

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490:55–60. doi: 10.1038/nature11450

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* 31, 2269–2275. doi: 10.1093/bioinformatics/btv165

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

Tang, Z. Z., Chen, G., Alekseyenko, A. V., and Li, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* 33, 1278–1285. doi: 10.1093/bioinformatics/btw804

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser.* 58, 267–288.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Virgin, H. W., and Todd, J. A. (2011). Metagenomics and personalized medicine. *Cell* 147, 44–56. doi: 10.1016/j.cell.2011.09.009

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14:508. doi: 10.1038/nrmicro.2016.83

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. *Gen. Med.* 8:56. doi: 10.1186/s13073-016-0302-3

Zhan, X., Epstein, M. P., and Ghosh, D. (2015). An adaptive genetic association test using double kernel machines. *Stat. Biosci.* 7, 262–281. doi: 10.1007/s12561-014-9116-2

Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. (2017a). A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics* 73, 1453–1463. doi: 10.1111/biom.12684

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017b). A small-sample multivariate kernel machine test for microbiome association studies. *Gen. Epidemiol.* 41, 210–220. doi: 10.1002/gepi.22030

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. *Gen. Epidemiol.* 42, 772–782. doi: 10.1002/gepi.22160

Zhang, X., Mallick, H., and Yi, N. (2016). Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *J. Bioinform. Genom.* 2:1. doi: 10.18454/jbg.2016.2.2.1

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* 21:895. doi: 10.1038/nm.3914

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. Gen.* 96, 797–807. doi: 10.1016/j.ajhg.2015.04.003

Zhao, N., Zhan, X., Guthrie, K. A., Mitchell, C. M., and Larson, J. (2018). Generalized hotelling's test for paired compositional data with application to human microbiome studies. *Gen. Epidemiol.* 42, 459–469.doi: 10.1002/gepi.22127

Conflict of Interest Statement: The authors declare that this study received funding from a National Institutes of Mental Health STTR award (R41 MH111347) to Quadrant Biosciences, Inc. Quadrant Biosciences was involved with study design, and data collection for the RNA sequencing results employed in this study's secondary data analysis (autism microbiome data). SH and FM serve on the scientific and medical advisory boards of Quadrant Biosciences Inc., and SH is a paid consultant for Quadrant Biosciences Inc.

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

METHODS
published: 16 May 2019
doi: 10.3389/fgene.2019.00458

A Distance-Based Kernel Association Test Based on the Generalized Linear Mixed Model for Correlated Microbiome Studies

*Hyunwook Koh[1], Yutong Li[2], Xiang Zhan[3], Jun Chen[4] and Ni Zhao[1]**

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [2] School of Physics, Peking University, Beijing, China, [3] Department of Public Health Sciences, Pennsylvania State University, Hershey, PA, United States, [4] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

Researchers have increasingly employed family-based or longitudinal study designs to survey the roles of the human microbiota on diverse host traits of interest (e. g., health/disease status, medical intervention, behavioral/environmental factor). Such study designs are useful to properly control for potential confounders or the sensitive changes in microbial composition and host traits. However, downstream data analysis is challenging because the measurements within clusters (e.g., families, subjects including repeated measures) tend to be correlated so that statistical methods based on the independence assumption cannot be used. For the correlated microbiome studies, a distance-based kernel association test based on the linear mixed model, namely, correlated sequence kernel association test (cSKAT), has recently been introduced. cSKAT models the microbial community using an ecological distance (e.g., Jaccard/Bray-Curtis dissimilarity, unique fraction distance), and then tests its association with a host trait. Similar to prior distance-based kernel association tests (e.g., microbiome regression-based kernel association test), the use of ecological distances gives a high power to cSKAT. However, cSKAT is limited to handling Gaussian traits [e.g., body mass index (BMI)] and a single chosen distance measure at a time. The power of cSKAT differs a lot by which distance measure is used. However, choosing an optimal distance measure is challenging because of the unknown nature of the true association. Here, we introduce a distance-based kernel association test based on the generalized linear mixed model (GLMM), namely, GLMM-MiRKAT, to handle diverse types of traits, such as Gaussian (e.g., BMI), Binomial (e.g., disease status, treatment/placebo) or Poisson (e.g., number of tumors/treatments) traits. We further propose a data-driven adaptive test of GLMM-MiRKAT, namely, aGLMM-MiRKAT, so as to avoid the need to choose the optimal distance measure. Our extensive simulations demonstrate that aGLMM-MiRKAT is robustly powerful while correctly controlling type I error rates. We apply aGLMM-MiRKAT to real familial and longitudinal microbiome data, where we discover significant disparity in microbial community composition by BMI status and the frequency of antibiotic use. In summary, aGLMM-MiRKAT is a useful analytical tool with its broad applicability to diverse types of traits, robust power and valid statistical inference.

Keywords: microbiome association studies, correlated microbiome studies, longitudinal microbiome studies, community-level association analysis, distance-based association analysis, adaptive association analysis

OPEN ACCESS

Edited by:
Himel Mallick,
Merck (United States), United States

Reviewed by:
Christine Burns Peterson,
University of Texas MD Anderson Cancer Center, United States
Ryan Sun,
Harvard University, United States
Michael B. Sohn,
University of Rochester, United States

***Correspondence:**
Ni Zhao
nzhao10@jhu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *08 February 2019*
Accepted: *30 April 2019*
Published: *16 May 2019*

Citation:
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

INTRODUCTION

The recent surge in next-generation sequencing technologies has dramatically advanced the human microbiome studies by enabling generic characterization of the microbes in the human body (Hamady and Knight, 2009; Caporaso et al., 2010; Thomas et al., 2012). As the sequencing technology evolves, researchers are able to obtain more accurate metagenomic information with lower cost at a faster speed. Various types of metagenomic information can be obtained by the sequencing platforms, such as microbial abundances and functional/metabolic expressions (Mallick et al., 2017). In this study, we focus on the data for the microbial abundance and phylogenetic information of the surrogate microbial species, known as, operational taxonomic units (OTUs). Furthermore, we focus on the microbiome association studies which test the disparity in microbial community (e.g., bacterial kingdom) composition by a host trait of interest (e.g., health/disease status, clinical intervention, behavioral/environmental factor) (Li, 2015). For example, recent studies have found disparity in microbial community composition for a variety of health/disease status [e.g., obesity (Arslan, 2014), type I diabetes (Zhang et al., 2018a), type II diabetes (Qin et al., 2012), human immunodeficiency virus (Bandera et al., 2018), inflammatory bowel disease (Knights et al., 2013; Borren et al., 2018), and cancers (Zitvogel et al., 2015)], medical interventions [e.g., administration of antibiotics (Zhang et al., 2018a)], and behavioral/environmental factors [e.g., diet, residence, smoking and birth mode (Charlson et al., 2010; Liu et al., 2017)].

Notably, researchers have increasingly employed family-based (Goodrich et al., 2014; Schloss et al., 2014) or longitudinal study designs (Yang et al., 2017; Zhang et al., 2018a). Such study designs are advantageous in properly controlling for potential confounders or the sensitive changes in microbial composition and host traits. That is, because family members share similar environmental/genetic factors (refer that monozygotic twins even have the same genetic background), the use of family controls can efficiently rule out some potential confounding factors. Moreover, because microbial composition and host traits can vary by time, repeated measurements over a lengthy follow-up period can ensure more reliable analysis outcomes. Examples for such correlated microbiome studies include the familial (Goodrich et al., 2014) and longitudinal (Zhang et al., 2018a) studies, the data of which we use for our real data applications (see Real data applications). Briefly, Goodrich et al. (2014) have collected stool samples from families with twins in the United Kingdom to assess the relationship between obesity and gut microbiota. Zhang et al. (2018a) longitudinally collected fecal, cecal, and ileal samples from non-obese diabetic mice to evaluate whether the intestinal microbiota altered by early-life antibiotic exposure affects maturation of innate immunity. The downstream data analysis for such studies is challenging because the measurements within clusters (e.g., families, subjects including repeated measures) tend to be correlated. We need to properly model the within-cluster correlation structure for valid statistical inferences. Besides, the unique features of the microbiome data (e.g., high-dimensionality, sparsity, and phylogenetic structure) need to be properly accounted for.

However, most of the current microbial community-level association tests [e.g., PERMANOVA (Anderson, 2001; McArdle and Anderson, 2001; Tang et al., 2016), MiRKAT (Zhao et al., 2015), MiSPU (Wu et al., 2016), OMiAT (Koh et al., 2017), aMiAD (Koh, 2018)] assume independent samples. Hence, they cannot be used for correlated microbiome studies. Zero-inflated Beta regression model (ZIBR) (Chen and Li, 2016) and negative Binomial mixed model (NBMM) (Zhang et al., 2017, 2018b) have recently been proposed for correlated microbiome studies. However, ZIBR and NBMM test individual microbial biomarkers (e.g., OTUs, taxa), not the microbial community as a whole. Hence, they are subject to a substantial loss of power after the requisite multiple testing correction. To our best knowledge, a remarkable community-level association test for correlated microbiome studies is the correlated sequence kernel association test (cSKAT) (Zhan et al., 2018). cSKAT is based on the linear mixed model (Laird and Ware, 1982), where the inherent random effect captures the within-cluster correlation of a host trait, and models the variance covariance structure of the microbial community based on an ecological distance, such as Jaccard dissimilarity (Jaccard, 1912), Bray-Curtis dissimilarity (Bray and Curtis, 1957) or unique fraction (UniFrac) distances (Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012). The use of ecological distances, which has also been widely adopted for many prior community-level association tests (Anderson, 2001; McArdle and Anderson, 2001; Zhao et al., 2015; Tang et al., 2016; Koh et al., 2017, 2018; Plantinga et al., 2017; Zhan et al., 2017), gives cSKAT a higher power than the ones based on non-ecological distances (Zhan et al., 2018). This is because the ecological distances are well-informed by properly modeling the microbial abundance and phylogenetic information (Jaccard, 1912; Bray and Curtis, 1957; Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012).

However, cSKAT has two major limitations. First, cSKAT is based on the linear mixed model (Laird and Ware, 1982). Hence, it is limited to handling Gaussian traits [e.g., body mass index (BMI)]. However, in practice, investigators can be interested in other trait types. Therefore, we introduce a distance-based kernel association test based on the generalized linear mixed model (GLMM), namely, GLMM-MiRKAT, to handle diverse types of traits, such as Gaussian (e.g., BMI), Binomial (e.g., disease status, treatment/placebo) or Poisson (e.g., number of tumors/treatments) traits. Second, cSKAT is limited to the item-by-item use of the ecological distances (i.e., the approach based on a single chosen ecological distance measure at a time). It is well-recognized in the microbiome research community that the power differs a lot by which distance measure is used, while it is also highly depending on the true underlying association pattern (Zhao et al., 2015; Koh et al., 2017, 2018). In practice, the true association pattern is usually unknown; hence, it is highly difficult to predict which distance measure performs best and choose a single optimal distance measure to use. The approach of individually testing multiple distances also requires multiple testing correction leading to a loss of power. Therefore, for a robustly high power, without the need to choose the

optimal distance measure, we propose a data-driven adaptive test of GLMM-MiRKAT, namely, aGLMM-MiRKAT. aGLMM-MiRKAT robustly adapts to diverse association patterns by jointly considering multiple candidate ecological distance measures. Jaccard dissimilarity (Jaccard, 1912), Bray-Curtis dissimilarity (Bray and Curtis, 1957), UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012) are included as the candidate ecological distance measures because of their well-known features and distinguished performances (details are addressed later) (Zhao et al., 2015). Through extensive simulation experiments, we estimate robustly high power with well-controlled type I error for aGLMM-MiRKAT.

The rest of the paper is organized as follows. (1) In Materials and Methods, we address methodological details. (2) In Simulation, we address extensive simulation experiments. (3) In Real data applications, we apply aGLMM-MiRKAT to real familial and longitudinal microbiome data sets, where we test the association of the microbial community composition with BMI and the frequency of antibiotic use, while making interesting testing attempts and interpretations. (4) In Discussion, we finish with discussion and concluding remarks.

MATERIALS AND METHODS

Notations and Models

We let y_{ij} denote a host trait of interest (e.g., health/disease status, medical intervention, behavioral/environmental factor) for the j-th measurement in the i-th cluster ($i = 1, \ldots, n, j = 1, \ldots, m_i$), z_{ijk} denote the abundance level of the k-th OTU among p OTUs in the microbial community ($k = 1, \cdots, p$), and x_{ijl} denote a covariate among q covariates (e.g., age, gender) that we want to adjust for ($l = 1, \ldots, q$). We also let N denote the total number of measurements (i.e., $N = \sum_{i=1}^{n} m_i$), $\boldsymbol{I_g}$ denote the g-th order identity matrix and $\boldsymbol{1_g}$ denote the $g \times 1$ vector of ones. Throughout the paper, we use non-bold lowercase letters for scalars, bold lowercase letters for vectors, and bold uppercase letters for matrices.

To relate the microbial community composition with a host trait adjusting for covariates, we consider a generalized linear mixed model (Breslow and Clayton, 1993) (Equation 1).

$$g(\mu_{ij}) = x_{ij}^T\boldsymbol{\alpha} + s_{ij}^T\boldsymbol{\upsilon_i} + h(z_{ij}), \tag{1}$$

where $g(\cdot)$ is a canonical link function (e.g., identity function for Gaussian traits, logistic function for Binomial traits, log function for Poisson traits) and $\mu_{ij} = E(y_{ij})$. $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_q)^T$ are fixed effects for the covariates $x_{ij} = (1, x_{ij1}, \ldots, x_{ijq})^T$. υ_i is the random effect for the pre-specified s_{ij} to account for the within-cluster correlation in responses (i.e., conditional on υ_i and $h(z_{ij})$, y_{ij} are independent with a diagonal variance-covariance matrix $\sigma_\varepsilon^2\boldsymbol{I}_{m_i}$). For example, when $s_{ij} = 1$, υ_i is the random intercept which is assumed to follow a normal distribution $N(0, \sigma_\gamma^2)$. When $s_{ij} = (1, t_{ij})^T$, where t_{ij} is the time point for the i-th cluster and j-th measurement, $\boldsymbol{\upsilon_i} = (\upsilon_{i1}, \upsilon_{i2})$ is the random intercept and slope which are assumed to follow normal distributions $\upsilon_{i1} \sim N(0, \sigma_{\gamma 1}^2)$ and $\upsilon_{i2} \sim N(0, \sigma_{\gamma 2}^2)$. Then, $\boldsymbol{\gamma_i} \equiv (s_{i1}\upsilon_i, \ldots, s_{im_i}\upsilon_i)^T$follows a normal distribution with mean zero and $m_i \times m_i$ variance-covariance matrix $\boldsymbol{\Sigma}_i$. The random effect υ_i is to capture the within-cluster correlation in responses, while $h(\cdot)$ is a function which features the microbiome effect.

Here, we are particularly interested in testing $\boldsymbol{H}_0$: $h(z_{ij}) = 0$ (i.e., no association between microbial composition and a host trait adjusting for covariates) and, notably, with different specifications for $h(z_{ij})$, we can characterize different association patterns between microbial composition and a host trait. One may specify $h(z_{ij})$ as a fixed effect using a linear or non-linear function for the OTUs. For example, we can specify $h(z_{ij}) = \varphi(z_{ij})^T\boldsymbol{\beta}$, where $\varphi(\cdot)$ is an element-wise transformation (e.g., identity or quadratic) function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are regression coefficients for the p OTUs, and then test H_0: $\boldsymbol{\beta} = \boldsymbol{0}$ using a p-degrees of freedom test. However, because of the high-dimensional nature of the data (i.e., $p >> n$) and, for example, the resulting issue of low-rank matrices, testing H_0: $\boldsymbol{\beta} = \boldsymbol{0}$ with fixed effects might be challenging or even impossible. Therefore, we apply the kernel trick (Cristianini and Shawe-Taylor, 2000) and specify $\delta_{ij} \equiv h(z_{ij}) = \sum_{i'=1}^{n}\sum_{j'=1}^{m_i}\omega_{ij}\kappa(z_{ij}, z_{j'}^{i'})$, where $\kappa(\cdot,\cdot)$ is a positive semi-definite kernel function which measures pairwise similarities in microbial composition, $z_{ij} = (z_{ij1}, \ldots, z_{ijp})^T$ is the $p \times 1$ vector for the p OTUs and ω_{ij}'s are coefficients; as such, $h(\cdot)$ lies in a reproducing kernel Hilbert space spanned by $\kappa(\cdot,\cdot)$. Then, via the connection between kernel machine regression and mixed effect models (Liu et al., 2007), $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{1m_1}, \ldots, \delta_{n1}, \ldots, \delta_{nm_n})^T$ is assumed to follow a distribution with mean zero and variance-covariance matrix $\tau\boldsymbol{K}$, where $\boldsymbol{\delta}$ is an $N \times 1$ vector, τ is the unknown variance component and $\boldsymbol{K}$ is an $N \times N$ pairwise similarity matrix. Then, we can perform a variance component test for H_0: $\tau = 0$ vs. H_1: $\tau > 0$ (Lin, 1997).

To address details on the kernel matrix $\boldsymbol{K}$ and the test statistic for H_0: $\tau = 0$, we first re-write the model (Equation 1) with matrix forms for all the measurements across all the clusters (Equation 2).

$$g(\mu) = \boldsymbol{X\alpha} + \boldsymbol{\gamma} + \boldsymbol{\delta}, \tag{2}$$

where $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{1m_1}, \ldots, \mu_{n1}, \ldots, \mu_{nm_n})^T$ is an $N \times 1$ vector, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_q)^T$ is an $(q+1) \times 1$ vector, $\boldsymbol{X} = (x_{11}, \ldots x_{1m_1}, \ldots, x_{n1}, \ldots, x_{nm_n})^T$ is an $N \times (q+1)$ matrix, $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_n})$ is an $N \times 1$ vector, and $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{1m_1}, \ldots, \delta_{n1}, \ldots, \delta_{nm_n})^T$ is an $N \times 1$ vector. Again, δ is assumed to follow a distribution with mean zero and variance-covariance matrix $\tau\boldsymbol{K}$. We further assume that the two random effects $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are independent as in (Lin, 1997). The kernel matrix $\boldsymbol{K}$ is an $N \times N$ pairwise similarity matrix which is converted from the use of an ecological distance (Zhao et al., 2015), such as Jaccard dissimilarity (Jaccard, 1912), Bray-Curtis dissimilarity (Bray and Curtis, 1957) or UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012), via (Equation 3).

$$K_{(h)} = -\frac{1}{2}\left(\boldsymbol{I_N} - \frac{\boldsymbol{1_N1_N^T}}{N}\right)\boldsymbol{D}_{(h)}^2\left(\boldsymbol{I_N} - \frac{\boldsymbol{1_N1_N^T}}{N}\right), \tag{3}$$

where $\boldsymbol{D}_{(h)}$ is the $N \times N$ pairwise distance matrix and $\boldsymbol{D}^2_{(h)}$ is its element-wise square matrix, where h is an index for a chosen measure among diverse ecological distances. This kernel matrix (Equation 3) externally models ecologically meaningful pairwise similarities (correlation) in microbial composition among all the measurements across all the clusters, where the block-diagonals (i.e., $\boldsymbol{K}_{(1,m_1),\,(1,m_1)}$, $\boldsymbol{K}_{(m_1+1,m_1+m_2),\,(m_1+1,m_1+m_2)}$, …, $\boldsymbol{K}_{(N-m_n+1,\,N),\,(N-m_n+1,\,N)}$) model the within-cluster similarities while the off-diagonals model the between-cluster similarities. The extent of OTU abundance and phylogenetic information is properly modulated by different ecological distance measures (Zhao et al., 2015).

GLMM-MiRKAT

While we will soon address the issue that the testing performance differs according to the choice of distance measure, we first introduce the variance component score statistic for a single chosen distance measure (i.e., item-by-item approach). Following (Lin, 1997), the variance component score statistic can be formulated with (Equation 4). Here, we construct the kernel matrix $K_{(h)}$ based on an ecological distance, and all the detailed derivation procedures are referred to (Lin, 1997).

$$\frac{\partial l(\boldsymbol{\alpha},\,\gamma,\,\tau)}{\partial \tau}\Big|_{\tau=0,\,\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_{\mathbf{0}},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{\mathbf{0}}} = \frac{1}{2}\left(\boldsymbol{y}^{*}-\boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\mathbf{0}}\right)^{T}\hat{\boldsymbol{V}}_{\mathbf{0}}^{-1}\boldsymbol{K}_{(h)}\hat{\boldsymbol{V}}_{\mathbf{0}}^{-1}(\boldsymbol{y}^{*}-\boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\mathbf{0}}) + tr(\hat{\boldsymbol{V}}_{\mathbf{0}}^{-1}\boldsymbol{K}_{(h)}), \tag{4}$$

where $\boldsymbol{y}^{*} = \boldsymbol{X}\hat{\boldsymbol{\alpha}}_{\mathbf{0}} + \hat{\boldsymbol{\gamma}}_{\mathbf{0}} + \hat{\boldsymbol{\Delta}}_{\mathbf{0}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\mathbf{0}})$ is the working vector and $\hat{\boldsymbol{V}}_{\mathbf{0}}^{-\mathbf{1}} = (\hat{\Sigma}_{\mathbf{0}} + \hat{\boldsymbol{W}}_{\mathbf{0}})^{-1}$. Here, $\hat{\boldsymbol{\Delta}}_{\mathbf{0}} = diag(g'(\hat{\boldsymbol{\mu}}_{\mathbf{0}}))$ (i.e., $\hat{\boldsymbol{\Delta}}_{\mathbf{0}} = \boldsymbol{I}_N$, $\hat{\boldsymbol{\Delta}}_{\mathbf{0}} = diag((\hat{\boldsymbol{\mu}}_{\mathbf{0}}(\mathbf{1}-\hat{\boldsymbol{\mu}}_{\mathbf{0}}))^{-1})$ and $\hat{\boldsymbol{\Delta}}_{\mathbf{0}} = diag(\hat{\boldsymbol{\mu}}_{\mathbf{0}}^{-\mathbf{1}})$ for Gaussian, Binomial, Poisson traits, respectively), $\hat{\boldsymbol{\Sigma}}_{\mathbf{0}} = diag(\hat{\boldsymbol{\Sigma}}_{\mathbf{1,0}}, \ldots, \hat{\Sigma}_{\boldsymbol{n,0}})$, and $\hat{\boldsymbol{W}}_{\mathbf{0}}$ is the dispersion parameter for the errors estimated as $\hat{\boldsymbol{W}}_{\mathbf{0}} = diag(var(\hat{\boldsymbol{\mu}}_{\mathbf{0}}), \ldots, var(\hat{\boldsymbol{\mu}}_{\mathbf{0}}))$ for Gaussian traits and $\hat{\boldsymbol{W}}_{\mathbf{0}} = \mathbf{I_N}$ for Binomial and Poisson traits, where $\hat{\boldsymbol{\alpha}}_{\mathbf{0}}$, $\hat{\boldsymbol{\gamma}}_{\mathbf{0}}$, $\hat{\boldsymbol{\mu}}_{\mathbf{0}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}$ are estimated under the null generalized linear mixed model by the restricted maximum likelihood estimation (REML) method (Harville, 1977) and $var(\cdot)$ is the variance function. This test statistic (Equation 4) is the penalized quasi-likelihood estimating equation in Breslow and Clayton (1993) and the variance component score statistic for testing random effects in Lin (1997) under the above model specifications. This is also the unadjusted variance component score statistic proposed for cSKAT which is based on the linear mixed model for Gaussian traits (Zhan et al., 2018). Similar test statistics have also been widely used for various family-based and longitudinal studies in genetics and neuroscience (Schifano et al., 2012; Chen et al., 2013; Zhang et al., 2014; Wang et al., 2017), while assuming different variance covariance structures and/or applying different weighting schema. Since our p-value computation is based on a permutation approach, the *scaling* (i.e., $\frac{1}{2}$) and *additive* [i.e., $tr(\hat{V}_0^{-1}K_{(h)})$] terms do not change the comparative ranks of the observed and null (i.e., permuted) statistic values (see *P*-value calculation). Hence, we use a reduced-form statistic (Equation 5).

$$Q_{(h)} = \left(\boldsymbol{y}^{*}-\boldsymbol{X}\hat{\alpha}_{\mathbf{0}}\right)^{T}\hat{\boldsymbol{V}}_{\mathbf{0}}^{-\mathbf{1}}\boldsymbol{K}_{(\boldsymbol{h})}\hat{\boldsymbol{V}}_{\mathbf{0}}^{-\mathbf{1}}(\boldsymbol{y}^{*}-\boldsymbol{X}\hat{\alpha}_{\mathbf{0}}) \tag{5}$$

aGLMM-MiRKAT

The testing performance depends on the choice of distance measure (Zhao et al., 2015). To explain, non-phylogeny-based distances, such as Jaccard (1912) and Bray and Curtis (1957) dissimilarities, measure the disparity only in abundance, while phylogeny-based distances, such as UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012), measure the disparity both in abundance and phylogeny. Hence, non-phylogeny-based distances are well-suited when associated OTUs have disparity in abundance, while phylogeny-based distances are well-suited when they have disparity both in abundance and phylogeny. Moreover, Jaccard dissimilarity and unweighted UniFrac distance are based on incidence information (i.e., presence/absence of OTUs), while Bray-Curtis dissimilarity and weighted UniFrac distance are based on full abundance information [refer that generalized UniFrac distance modulates the intensity of abundance information between unweighted and weighted UniFrac distances by its parameter θ (Chen et al., 2012)]. Hence, Jaccard dissimilarity and unweighted UniFrac distance are well-suited when associated OTUs are rare in abundance in the sense that prevalent OTUs are likely to exist in all samples, while Bray-Curtis dissimilarity and weighted UniFrac distance are well-suited when they are rich in abundance. However, prior knowledge about the true association pattern is usually absent in reality. Hence, it is highly challenging to choose a single optimal distance measure to use. For a robustly high performance throughout various (but unknown) association scenarios, we propose aGLMM-MiRKAT which is based on the test statistic of the minimum p-value from multiple item-by-item GLMM-MiRKAT analyses (Equation 6).

$$T_{aGLMMMiKAT} = \min_{h\in\Gamma} P_{(h)}, \tag{6}$$

where h is an index for a distance in a set of candidate ecological distances (Γ), where Γ = {Jaccard dissimilarity, Bray-Curtis dissimilarity, Unweighted UniFrac distance, Generalized UniFrac distance ($\theta = 0.5$), Weighted UniFrac distance}. Obviously, we do not report the genuine minimum p-value (i.e., $T_{aGLMMMiKAT}$) as it is. Instead, $T_{aGLMMMiKAT}$ (Equation 6) is the test statistic of aGLMM-MiRKAT, and we estimate the p-value for aGLMM-MiRKAT ($P_{aGLMMMiKAT}$) using a permutation approach (see *P*-value calculation). Our extensive simulations reveal that aGLMM-MiRKAT maintains high power throughout all surveyed association scenarios, while the item-by-item GLMM-MiRKAT analyses are limitedly powerful only for some association scenarios. Further details are addressed in the Simulation section.

P-value Calculation

We calculate the p-values for the item-by-item GLMM-MiRKAT tests and aGLMM-MiRKAT using a permutation approach. Our permutation approach is semi-parametric as we fit the null model $g(\hat{\mu}_0) = X\hat{\alpha}_0 + \hat{\gamma}_0$ (Equation 2) (excluding the microbiome portion) parametrically, and then draw the empirical null distribution of the test statistic (Equations 5, 6) through permutations non-parametrically. In this way, we can estimate the p-values without making distributional assumptions for the

microbiome portion. Moreover, we do block permutations to account for any potential mis-specified within-cluster correlation structure based on the procedures in (Winkler et al., 2015). To be specific, for the random intercept model [i.e., $r_{ij} = 1$ (Equation 1)], we permute (1) the whole clusters (only the exchangeable clusters which have the same number of measurements) and (2) the measurements within each cluster, simultaneously. For the random slope model [i.e., $r_{ij} = \left(1,\ t_{ij}\right)^T$ (Equation 1)], we permute only the whole clusters (the exchangeable clusters which have the same number of measurements and the same time points). The detailed procedures for our permutation approach can be found in **S1. Computational algorithm**.

RESULTS

Simulation

Simulation Designs

Our simulation designs are based on prior studies (Zhao et al., 2015; Koh et al., 2017; Zhan et al., 2018), but here we conduct more extensive simulation experiments for diverse trait types with different within-cluster correlation structures. In particular, we simulated the data for Gaussian, Binomial and Poisson traits, respectively, based on the following generalized linear mixed models.

$$\begin{aligned} y_{ij} &= 0.5\times scale(\boldsymbol{x_{i1}} + x_{ij2}) \\ &\quad + \beta\times scale(\textstyle\sum_{a\in\mathcal{A}} \boldsymbol{z_{ija}}) + s_{ij}^T \upsilon_i + \epsilon_{ij} \\ logit(E(y_{ij}=1)) &= 0.5\times scale(\boldsymbol{x_{i1}} + x_{ij2}) \\ &\quad + \beta\times scale(\textstyle\sum_{a\in\mathcal{A}} \boldsymbol{z_{ija}}) + s_{ij}^T \upsilon_i \\ log(E(y_{ij})) &= 0.5\times scale(\boldsymbol{x_{i1}} + x_{ij2}) \\ &\quad + \beta\times scale(\textstyle\sum_{a\in\mathcal{A}} \boldsymbol{z_{ija}}) + s_{ij}^T \upsilon_i \end{aligned}$$

In these equations, $\boldsymbol{x_{i1}}$ is a cluster-specific (e.g., gender) covariate generated from the Bernoulli distribution with success probability 0.5, and x_{ij2} is a non-cluster-specific (e.g., time-varying) covariate generated from $0.5 \times scale(\sum_{a\in\mathcal{A}} \boldsymbol{z_{ija}}) + N(0, 1)$. Note that, x_{ij2} is a confounder as it is associated with both of the microbial composition and host trait. $\mathcal{A}$ is a set of associated OTUs among the total p OTUs in the community, and $\boldsymbol{z_{ija}}$ is the *a-th* OTU in $\mathcal{A}$. β is a regression coefficient for the OTUs in $\mathcal{A}$. *scale* is the standardization function to have mean zero and standard deviation one. υ_i is the random effect for the pre-specified s_{ij}, and ε_{ij} are errors generated from $N(0, 1)$. We investigate small ($n = 20$) and moderate ($n = 50$) numbers of clusters, respectively, while assigning two, three and four measurements, respectively, into each one third of the clusters (i.e., when $n = 20$, $m_i = 2$ for $i = 1, \ldots, 7$, $m_i = 4$ for $i = 8, \ldots, 14$ and $m_i = 3$ for $i = 15, \ldots, 20$; when $n = 50$, $m_i = 2$ for $i = 1, \ldots, 17$, $m_i = 3$ for $i = 18, \ldots, 34$ and $m_i = 4$ for $i = 35, \ldots, 50$). This is to mimic (possibly) unbalanced numbers of measurements across clusters. As before, we let $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, $k = 1, \ldots, p$ and $l = 1, \ldots, q$. For the random effect ν_i, we generate (1) random intercepts and (2) random intercepts and slopes, respectively, as follows. For the random intercepts (i.e., $s_{ij} = 1$), we generate ν_i from $N(0, \sigma_\gamma^2)$, while setting $\sigma_\gamma^2 = \frac{1}{2}$, 1 and $\frac{3}{2}$, respectively, to investigate different within-cluster correlations, that is, $\rho_{j\neq j'} = \sigma_\gamma^2/(\sigma_\gamma^2+\sigma_\varepsilon^2) = \frac{1}{3}$, $\frac{1}{2}$ and $\frac{3}{5}$. For the random intercepts and slopes (i.e., $\boldsymbol{s_{ij}} = \left(1,\ j\right)^T$), we generate ν_{i1} and ν_{i2} from $N(0, \sigma_\gamma^2)$, while setting $\sigma_\gamma^2 = \frac{1}{2}$, 1 and $\frac{3}{2}$, respectively and $t_{ij} = j$, to investigate different within-cluster correlations, that is, $\rho_{j\neq j'} = \sigma_\gamma^2/(\sigma_\gamma^2+\sigma_\varepsilon^2)$ $= \frac{(1+j^2)}{(j^2+3)}$, $\frac{(1+j^2)}{(j^2+2)}$ and $\frac{(1+j^2)}{(j^2+\frac{5}{3})}$.

For the OTUs in the community, we first estimated proportional means and a dispersion parameter for 856 OTUs (i.e., $p = 856$) in the bacterial kingdom from the real respiratory-tract microbiome data (Charlson et al., 2010). Then, OTU counts for each measurement per cluster (i.e., Z_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$) were generated from the Dirichlet-multinomial distribution (Mosimann, 1962) with the pre-specified parameter values of the estimated proportional means and dispersion. The total reads for each measurement were set to be 10,000. To reflect possible within-cluster relatedness among microbial communities, we updated the second and third measurements of microbial community using a random perturbation function: $Z_{ij} = \frac{1}{2}(Z_{i(j-1)} + Z_{ij})$ for $j = 2, \ldots, m_i$.

To estimate empirical type I error rates, we set $\beta = 0$. To estimate statistical powers, we set $\beta = 1$, while selecting a set of associated OTUs ($\mathcal{A}$) by four different association scenarios as in Koh et al. (2017, 2018) and Koh (2018) (1) 50 random OTUs among the OTUs in lower half of abundance, (2) 50 random OTUs, (3) 50 random OTUs among the OTUs in upper half of abundance, and (4) OTUs in a cluster among 10 clusters partitioned by the partition around medoids (PAM) algorithm (Reynolds et al., 2006) based on OTUs' cophenetic distances (Sneath et al., 1975), respectively. The first three scenarios mimic the situations when associated OTUs are rare, medium and abundant, respectively, while the fourth scenario mimics the situation when they are close in phylogeny. For the fourth scenario, we randomized the selection of an associated cluster among the 10 clusters to avoid arbitrary cluster selection. To estimate empirical type I error rates, we conducted 30,000 replicates for each combination of the model, sample size and correlation structure. To estimate statistical powers, we conducted 10,000 replicates for each combination of the model, sample size, correlation structure and association scenario.

Model fitting

We fit the random intercept model (i.e., $s_{ij} = 1$) when the random intercepts are generated, and we fit the random slope model (i.e., $\boldsymbol{s_{ij}} = \left(1,\ j\right)^T$) when the random intercepts and slopes are generated, while including the two covariates and all the 856 OTUs in the community.

Simulation Outcomes

Type I error

We estimate well-controlled empirical type I error rates at the significance level of 0.05 for any item-by-item GLMM-MiRKAT or aGLMM-MiRKAT test, for any type of traits (i.e., Gaussian, Binomial and Poisson traits), for both small ($n = 20$) and moderate ($n = 50$) numbers of clusters, for any imposed within-cluster correlation, and for both random intercept (**Table 1**) and slope models (**Table 2**). However, we

TABLE 1 | Estimated type I error rates at the significance level of 5% for GLMM-MiRKAT/aGLMM-MiRKAT based on the random intercept model with Gaussian, Binomial or Poisson responses (Unit: %).

	$n = 20$			$n = 50$		
$\rho_{j\neq j'}$	L	M	H	L	M	H
Gaussian						
K_J	5.06	4.89	5.12	5.08	5.06	4.98
K_{BC}	4.78	4.80	4.85	4.83	4.86	4.73
K_U	5.07	4.96	5.04	5.19	5.05	5.06
$K_{0.5}$	5.03	4.83	4.94	5.15	4.95	4.74
K_W	4.97	5.00	4.91	4.75	4.73	4.54
adaptive	4.89	4.74	4.74	4.92	4.79	4.73
Binomial						
K_J	5.08	4.93	4.91	5.00	5.13	4.88
K_{BC}	4.98	4.95	4.92	5.29	5.00	4.96
K_U	5.09	5.04	5.00	5.08	5.19	4.74
$K_{0.5}$	5.05	4.88	4.89	5.03	5.13	5.12
K_W	4.92	4.89	5.04	5.11	4.90	5.11
adaptive	4.87	4.90	4.89	5.06	4.99	4.92
Poisson						
K_J	4.98	4.93	5.11	4.95	5.17	5.06
K_{BC}	5.04	5.03	4.69	5.01	4.95	5.03
K_U	5.07	4.85	5.16	4.95	5.17	5.06
$K_{0.5}$	5.10	4.92	4.85	4.97	4.95	5.02
K_W	5.11	4.87	4.64	5.03	5.09	4.90
adaptive	4.96	4.91	4.83	4.95	5.00	5.07

K_J: Jaccard dissimilarity; K_{BC}: Bray-Curtis dissimilarity; K_U: Unweighted UniFrac distance; $K_{0.5}$: Generalized UniFrac distance (θ =0.5); K_W: Weighted UniFrac distance; adaptive: adaptive GLMM-MiRKAT (aGLMM-MiRKAT). L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{5}$); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{2}$); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{3}{5}$).

TABLE 2 | Estimated type I error rates at the significance level of 5% for GLMM-MiRKAT/aGLMM-MiRKAT based on the random slope model with Gaussian, Binomial or Poisson responses (Unit: %).

	$n = 20$			$n = 50$		
$\rho_{j\neq j'}$	L	M	H	L	M	H
Gaussian						
K_J	5.10	4.96	5.12	4.87	4.98	5.04
K_{BC}	5.11	4.89	4.97	5.10	4.88	5.03
K_U	5.03	4.95	5.13	5.03	5.03	5.10
$K_{0.5}$	5.07	4.91	4.90	4.89	4.91	5.09
K_W	4.96	4.95	4.87	4.83	5.03	5.01
adaptive	4.97	4.94	5.01	4.94	4.86	5.04
Binomial						
K_J	5.08	4.80	5.01	5.09	5.02	4.83
K_{BC}	4.93	4.94	5.1	4.89	5.02	4.88
K_U	5.04	4.99	5.04	5.07	5.40	4.83
$K_{0.5}$	5.02	4.97	4.84	5.00	5.08	4.96
K_W	4.89	5.07	5.02	4.96	5.08	4.85
adaptive	4.99	4.94	4.85	4.86	5.11	4.82
Poisson						
K_J	5.01	4.98	4.76	4.93	5.10	4.90
K_{BC}	5.16	4.76	5.02	5.03	5.03	5.02
K_U	4.90	5.06	4.92	5.09	5.19	4.93
$K_{0.5}$	5.14	4.87	5.10	4.85	4.88	5.10
K_W	5.12	4.82	5.28	4.86	5.06	5.18
adaptive	5.05	4.70	4.88	5.00	4.94	4.78

K_J: Jaccard dissimilarity; K_{BC}: Bray-Curtis dissimilarity; K_U: Unweighted UniFrac distance; $K_{0.5}$: Generalized UniFrac distance (θ =0.5); K_W: Weighted UniFrac distance; adaptive: adaptive GLMM-MiRKAT (aGLMM-MiRKAT). L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{(1+l^2)}{(l^2+3)}$); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{(1+l^2)}{(l^2+2)}$); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{(1+l^2)}{(l^2+\frac{5}{3})}$).

estimate inflated empirical type I error rates (>0.05) for the prior microbial community-level association tests, OMiRKAT (Zhao et al., 2015), aMiSPU (Wu et al., 2016), OMiAT (Koh et al., 2017), and aMiAD (Koh, 2018) (**Table 3**). This is because these tests treat all the measurements across all the clusters as independent samples in an exaggerated manner. We also observe in general that the higher the within-cluster correlation, the greater the type I error inflation (**Table 3**), as explained by the higher the within-cluster correlation, the smaller the effective sample size.

Power

We estimate in general that the moderate number of clusters (n =50) (**Figures 1**, **2**) is more powerful than the small number of clusters (n = 20) (**Figures S1**, **S2**), yet we observe the same comparative powers among different GLMM-MiRKAT analyses for the small (n = 20) and moderate (n = 50) number of clusters. Thus, to save space, the power outcomes for the small (n = 20) number of clusters are placed in (**Figures S1**,**S2**).

We estimate in general that the Gaussian models (**Figures 1A–C**, **2A–C**) are more powerful than the Binomial (**Figures 1D–F**, **2D–F**) and Poisson (**Figures 1G–I**, **2G–I**) models, where the Binomial models are the least powerful. This is because the continuous traits are better informed than the discrete traits, but not because our methods better suit the Gaussian models. We also observe in general that the higher the within-cluster correlation, the lower the power (i.e., **Figures 1A,D,G**, **2A,D,G** > **Figures 1B,E,H**, **2B,E,H** > **Figures 1C,F,I**, **2C,F,I**), as explained by the higher the within-cluster correlation, the smaller the effective sample size. We observe similar comparative powers among different GLMM-MiRKAT analyses across Gaussian, Binomial and Poisson models for both of the random intercept (**Figure 1**) and slope (**Figure 2**) models. We address the detailed description on the comparative powers below.

GLMM-MiRKAT using Jaccard dissimilarity or unweighted UniFrac distance is more powerful in the first scenario when associated OTUs are rare in abundance (**Figures 1**, **2**: P1), while GLMM-MiRKAT using Bray-Curtis dissimilarity or weighted UniFrac distance is relatively more powerful in the second and third scenarios when associated OTUs are mid-abundant and abundant (**Figures 1**, **2**: P2-P3), as expected by their distinct weighting schema. GLMM-MiRKAT using weighted UniFrac distance or generalized UniFrac distance is more powerful in the fourth scenario when associated OTUs are close in

TABLE 3 | Estimated type I error rates at the significance level of 5% for the prior microbial community-level association tests, OMiRKAT, aMiSPU, OMiAT, and aMiAD, for the clustered microbiome data (Unit: %).

Random intercepts						
	$n = 20$			$n = 50$		
$\rho_{j\neq j'}$	**L**	**M**	**H**	**L**	**M**	**H**
Gaussian						
OMiRKAT	24.36	79.89	97.44	37.98	96.61	99.96
aMiSPU	14.64	52.5	80.78	20.47	75.69	95.65
OMiAT	22.13	79.27	97.77	40.63	98.65	99.97
aMiAD	5.70	6.79	8.22	6.11	7.39	8.82
Binomial						
OMiRKAT	7.12	20.19	41.40	9.35	30.02	62.19
aMiSPU	6.17	12.32	24.13	6.88	16.18	34.86
OMiAT	6.87	18.54	39.62	9.09	33.68	71.1
aMiAD	5.41	5.71	6.31	5.64	5.98	6.62
Random intercepts and slopes						
Gaussian						
OMiRKAT	81.86	99.27	99.89	97.53	99.92	99.94
aMiSPU	72.20	96.42	98.58	92.87	99.88	99.98
OMiAT	81.31	99.41	99.91	98.70	99.93	99.97
aMiAD	8.59	10.68	11.57	8.51	10.24	10.58
Binomial						
OMiRKAT	23.98	63.69	84.53	36.73	86.82	97.98
aMiSPU	15.87	42.33	62.83	21.83	63.68	84.62
OMiAT	22.64	63.08	85.10	40.63	93.27	99.49
aMiAD	6.15	7.30	8.35	6.20	7.45	8.24

L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{3}$ for the random intercepts, $\rho_{j\neq j'} = \frac{(1+j^2)}{(j^2+3)}$ for the random intercepts and slopes); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{2}$ for the random intercepts, $\rho_{j\neq j'} = \frac{(1+j^2)}{(j^2+2)}$ for the random intercepts and slopes); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{3}{5}$ for the random intercepts, $\rho_{j\neq j'} = \frac{(1+j^2)}{(j^2+\frac{5}{3})}$ for the random intercepts and slopes).

phylogeny (**Figures 1**, **2**: P4), where GLMM-MiRKAT using Jaccard dissimilarity or Bray-Curtis dissimilarity is less powerful (**Figures 1**, **2**: P4), as expected by their use or non-use of phylogenetic information. Notably, none of the item-by-item GLMM-MiRKAT analyses are consistently powerful throughout all different association scenarios (i.e., they are powerful for some scenarios to which they are well-suited, but they are under-powered for the other scenarios to which they are not well-suited) (**Figures 1**, **2**). On the contrary, we estimate that the adaptive test of GLMM-MiRKAT, aGLMM-MiRKAT, is robustly powerful (closely reaching the highest power among the item-by-item GLMM-MiRKAT analyses) throughout all different association scenarios (**Figures 1**,**2**).

We additionally compare aGLMM-MiRKAT with the item-by-item cSKAT analyses for the random intercept Gaussian models as cSKAT can handle only the Gaussian traits based on the random intercept model (Zhan et al., 2018). Similar to the previous item-by-item GLMM-MiRKAT analysis outcomes, none of the item-by-item cSKAT analyses are consistently powerful throughout all different association scenarios (i.e., they are powerful for some scenarios to which they are well-suited, but they are under-powered for the other scenarios to which they are not well-suited) (**Figure 3**). Here again, we observe that aGLMM-MiRKAT maintains a high power throughout all different scenarios (**Figure 3**).

Real Data Applications

A Family-Based Study on the Association Between Obesity and Gut Microbiota

Goodrich et al. (2014) have collected fecal samples from the United Kingdom twin population to study the roles of host genetics on gut microbiome, while addressing a breadth of associations between obesity indices and gut microbiota. Here, we analyze a small portion the original data to evaluate the association between BMI and microbial community composition. The raw sequence data are publicly available in the European Bioinformatics Institute (EBI) repository (Assess codes: ERP006339 and ERP006342). We processed them using the QIIME pipeline (Caporaso et al., 2010) with open reference-based OTU picking by targeting the V4 region of the 16S ribosomal RNA (rRNA) gene, and quantified OTUs at the 97% sequence similarity level and constructed a phylogenetic tree. Among the total of 1,024 measurements from 536 families, we focused on monozygotic twins. After excluding measurements with low sequencing depth (i.e., <10,000 total reads), 311 measurements from 145 families were included in our analysis. The data originally include 7,365 OTUs, but we removed OTUs with average relative abundance $< 10^{-5}$, and then the data were rarefied to control unequal library sizes (Weiss et al., 2017); as such, 2,128 OTUs were included in our analysis.

We first visually check with principle coordinate analysis (PCoA) plots based on each distance measure to see if there is any disparity in microbial composition by BMI categories [i.e., under-weighted: BMI $(\frac{kg}{m^2}) <$ 18.5; normal: 18.5 $\leq$ BMI $(\frac{kg}{m^2}) <$ 25; over-weighted: 25 $\leq$ BMI $(\frac{kg}{m^2}) <$ 30; obese: 30 $\leq$ BMI $(\frac{kg}{m^2})$] (**Figure 4**). It is not very clear in the visual inspection if there is any significant separation by BMI categories, and we observe the smallest separation based on weighted UniFrac distance (**Figure 4**).

We fitted GLMM-MiRKAT with random intercepts for BMI in continuous scale (Gaussian traits) adjusting for age. GLMM-MiRKAT using Jaccard dissimilarity (p-value: <0.001), Bray-Curtis dissimilarity (p-value: <0.001), unweighted UniFrac distance (p-value: <0.001) or generalized UniFrac distance ($\theta = 0.5$) (p-value: 0.005) estimates significant association between BMI and microbial composition, while GLMM-MiRKAT using weighted UniFrac distance (p-value: 0.157) does not. This matches with our visual inspection of the smallest separation for the weighted UniFrac distance (**Figure 4**). This also indicates that the item-by-item GLMM-MiRKAT analyses are considerably sensitive to the choice of distance measure.

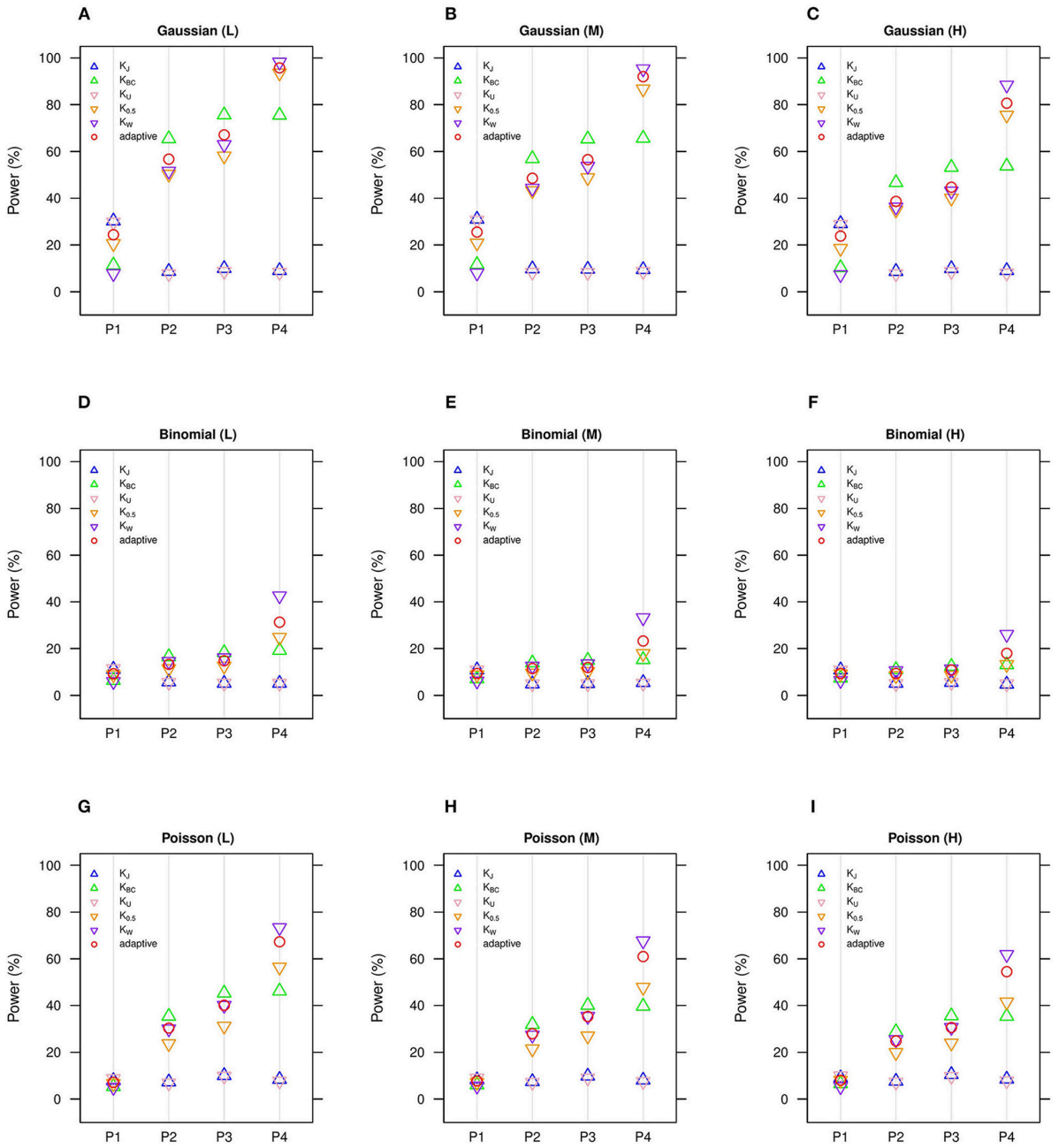


FIGURE 1 | Estimated statistical powers for GLMM-MiRKAT/aGLMM-MiRKAT based on the random intercept model with Gaussian, Binomial or Poisson responses ($n = 50$) (Unit: %). L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{3}$); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{2}$); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{3}{5}$).$K_J$: Jaccard dissimilarity; K_{BC}: Bray-Curtis dissimilarity; K_U: Unweighted UniFrac distance; $K_{0.5}$: Generalized UniFrac distance ($\theta = 0.5$); K_W: Weighted UniFrac distance; *adaptive*: adaptive GLMM-MiRKAT (aGLMM-MiRKAT). P1, P2, P3, and P4 represent the four different association scenarios: P1. $\mathcal{A} = \{50$ random OTUs in lower half of abundance$\}$; P2. $\mathcal{A} = \{50$ random OTUs$\}$; P3. $\mathcal{A} = \{50$ random OTUs in upper half of abundance$\}$; P4. $\mathcal{A} = \{$A random cluster among 10 clusters partitioned by PAM$\}$. **(A)** Gaussian (L); **(B)** Gaussian (M); **(C)** Gaussian (H); **(D)** Binomial (L); **(E)** Binomial (M); **(F)** Binomial (H); **(G)**. Poisson (L); **(H)** Poisson (M); **(I)**. Poisson (H).

aGLMM-MiRKAT estimates the significant association (p-value: <0.001).

For another demonstration, we fitted GLMM-MiRKAT with random intercepts for BMI in binary scale (Binomial traits) adjusting for age, comparing the normal and obese populations (i.e., 140 measurements from 85 families in the normal vs. 63 measurements from 41 families in the obese). However, we could not find any significant association by any item-by-item [i.e., Jaccard dissimilarity (p-value: 0.354), Bray-Curtis dissimilarity (p-value: 0.107), unweighted UniFrac distance (p-value: 0.336), generalized UniFrac distance ($\theta = 0.5$) (p-value: 0.231), weighted UniFrac distance (p-value: 0.333)] or adaptive

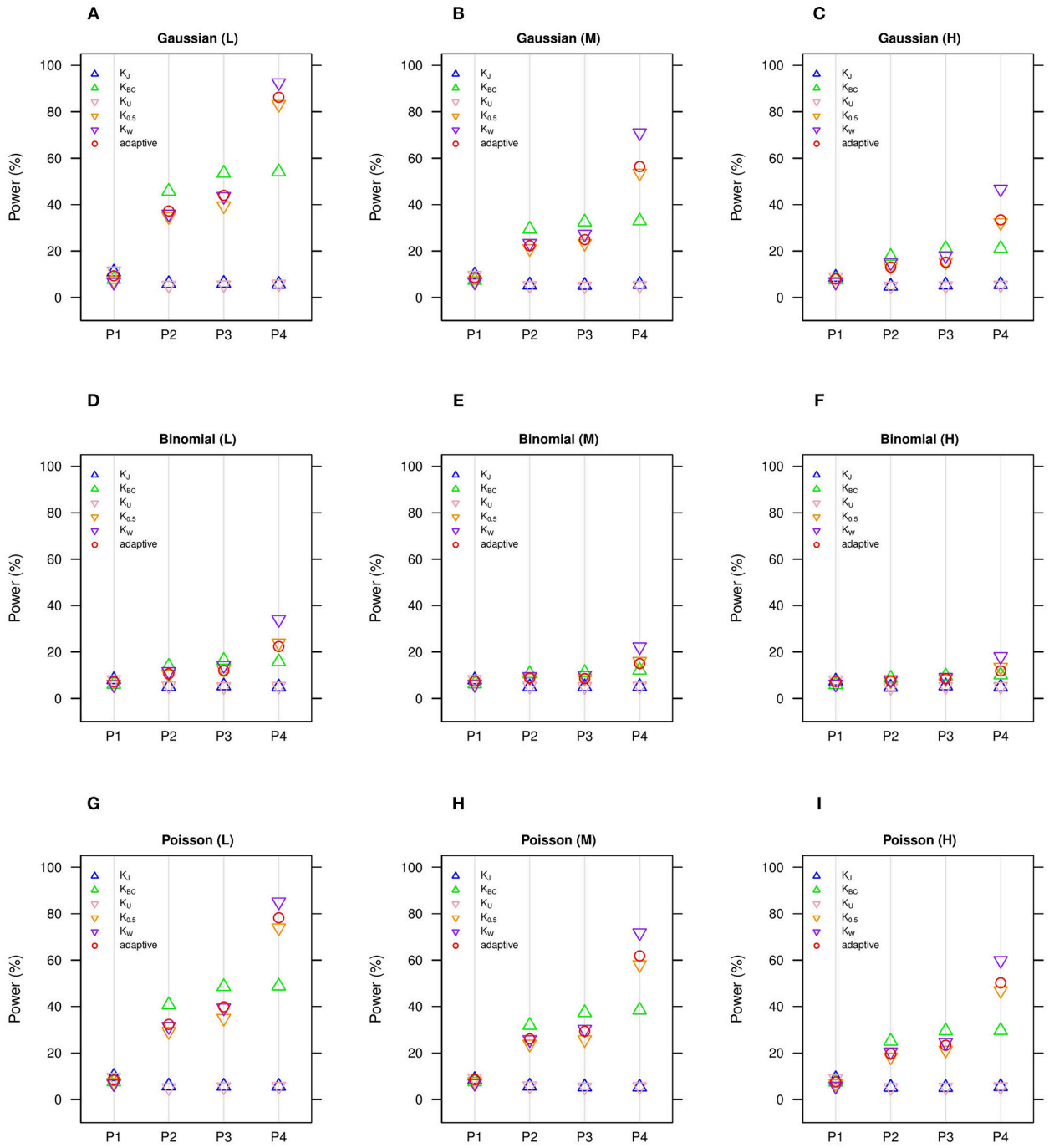


FIGURE 2 | Estimated statistical powers for GLMM-MiRKAT/aGLMM-MiRKAT based on the random slope model with Gaussian, Binomial or Poisson responses (n = 50) (Unit: %). L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{3}$); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{2}$); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{3}{5}$). K_J: Jaccard dissimilarity; K_{BC}: Bray-Curtis dissimilarity; K_U: Unweighted UniFrac distance; $K_{0.5}$: Generalized UniFrac distance ($\theta = 0.5$); K_W: Weighted UniFrac distance; *adaptive*: adaptive GLMM-MiRKAT (aGLMM-MiRKAT). P1, P2, P3, and P4 represent the four different association scenarios: P1. $\mathcal{A}$ = {50 random OTUs in lower half of abundance}; P2. $\mathcal{A}$ = {50 random OTUs}; P3. $\mathcal{A}$ = {50 random OTUs in upper half of abundance}; P4. $\mathcal{A}$ = {A random cluster among 10 clusters partitioned by PAM}. **(A)** Gaussian (L); **(B)** Gaussian (M); **(C)** Gaussian (H); **(D)** Binomial (L); **(E)** Binomial (M); **(F)** Binomial (H); **(G)** Poisson (L); **(H)** Poisson (M); **(I)** Poisson (H).

[i.e., aGLMM-MiRKAT (p-value: 0.253)] analysis. This power loss, of course, is related to the reduced sample size in the selected comparison. This may also indicate that BMI in continuous scale is better informed than BMI in binary scale, which matches with our simulation result, where the Gaussian models are more powerful than the Binomial models (**Figures 1**,**2**).

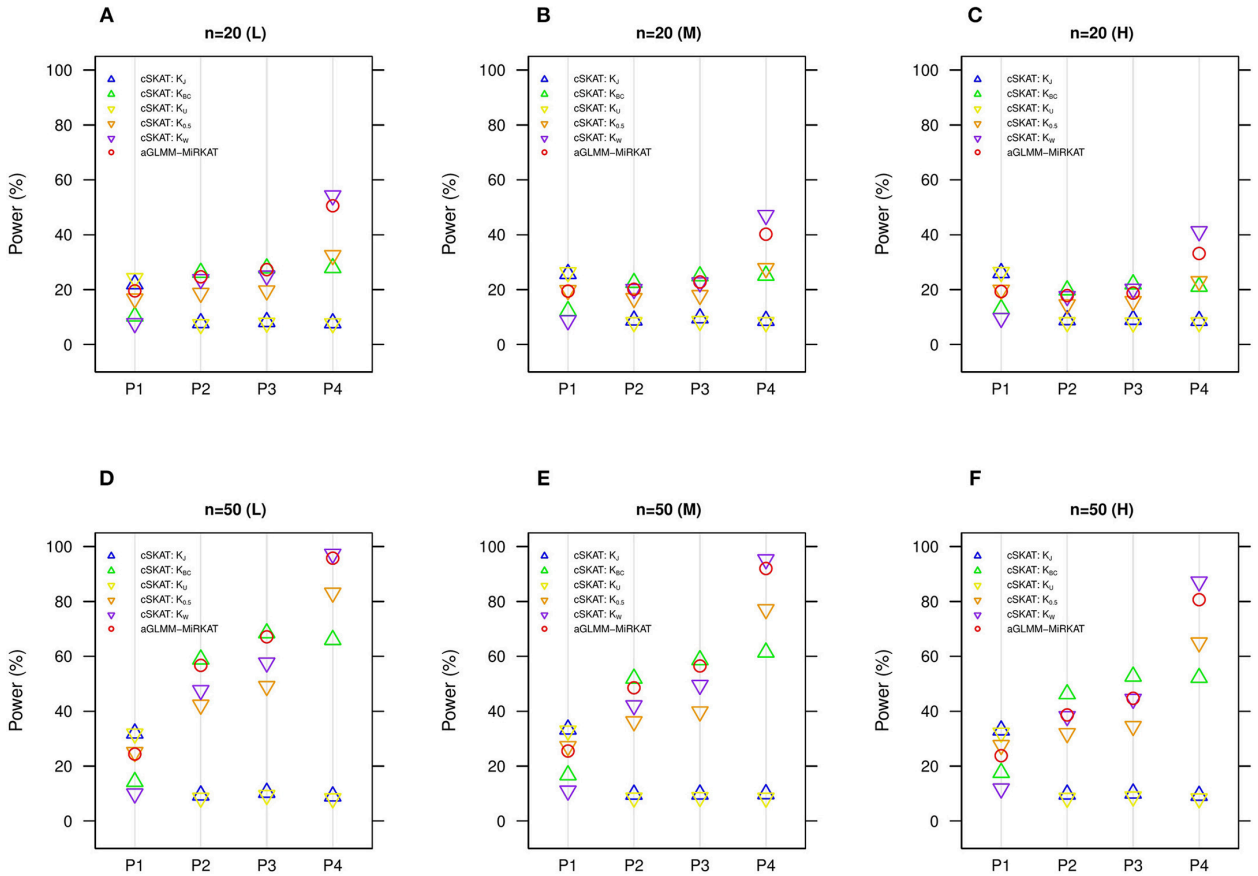


FIGURE 3 | Estimated statistical powers for the item-by-item cSKAT tests and aGLMM-MiRKAT based on the random intercept model with Gaussian responses (n=50) (Unit: %). L: low within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{3}$); M: medium within-cluster correlation ($\rho_{j\neq j'} = \frac{1}{2}$); H: high within-cluster correlation ($\rho_{j\neq j'} = \frac{3}{5}$).$K_J$: cSKAT for Jaccard dissimilarity; K_{BC} : cSKAT for Bray-Curtis dissimilarity; K_U : cSKAT for Unweighted UniFrac distance; $K_{0.5}$: cSKAT for Generalized UniFrac distance (θ =0.5); K_W: cSKAT for Weighted UniFrac distance; *adaptive*: adaptive GLMM-MiRKAT (aGLMM-MiRKAT). P1, P2, P3, and P4 represent the four different association scenarios: P1. $\mathcal{A}$ = {50 random OTUs in lower half of abundance}; P2. $\mathcal{A}$ = {50 random OTUs}; P3. $\mathcal{A}$ = {50 random OTUs in upper half of abundance}; P4. $\mathcal{A}$ = {A random cluster among 10 clusters partitioned by PAM}. **(A)** n = 20 (L); **(B)** n = 20 (M); **(C)** n = 20 (H); **(D)** n = 50 (L); **(E)** n = 50 (M); **(F)** n = 50 (H).

A Longitudinal Study on the Association Between the Frequency of Antibiotic Use and Gut Microbiota

Zhang et al. (2018a) collected fecal, cecal and ileal samples from non-obese diabetic mice for microbiome profiling studies based on a longitudinal study design to evaluate if the intestinal microbiota altered by early-life antibiotic exposure affects maturation of innate immunity. The raw sequence data are publicly available in the Qiita database (Identifier: 11242). We processed them using the QIIME pipeline (Caporaso et al., 2010) with open reference-based OTU picking by targeting the V4 region of the 16S rRNA gene, and quantified OTUs at the 97% sequence similarity level and constructed a phylogenetic tree. The original study (Zhang et al., 2018a) contains enormous amount of data for a number of sub-studies, but, for a demonstration of our proposed method, we only analyze a small portion of the data. To be specific, we focused on fecal samples to evaluate the disparity in microbial community composition by the frequency of antibiotic use (i.e., 0, 1, 2, and 3 course(s) of antibiotic use). After excluding measurements with low sequencing depth (i.e., <10,000 total reads), 229 measurements from 87 mice were included in our analysis. The study design is longitudinal and unbalanced in that each mouse has different numbers of repeated measurements: 61 mice have three measurements, 20 mice have two measurements and 6 mice have one measurement through different time points. Among the total of 229 measurements, 120 have had no antibiotic use, 43 have had one course of antibiotic use, 26 have had two courses of antibiotic use, and 40 have had three courses of antibiotic use.

Here, we first visually check with the PCoA plots based on each distance measure to see if there is any disparity in microbial composition by different numbers of antibiotic use (**Figure 5**). We observe a very clear visual separation, especially from no antibiotic use group to at least one course of antibiotic use group, based on any distance measures (**Figure 5**).

We fitted GLMM-MiRKAT with random intercepts for the number of antibiotic use (Poisson traits) (i.e.,

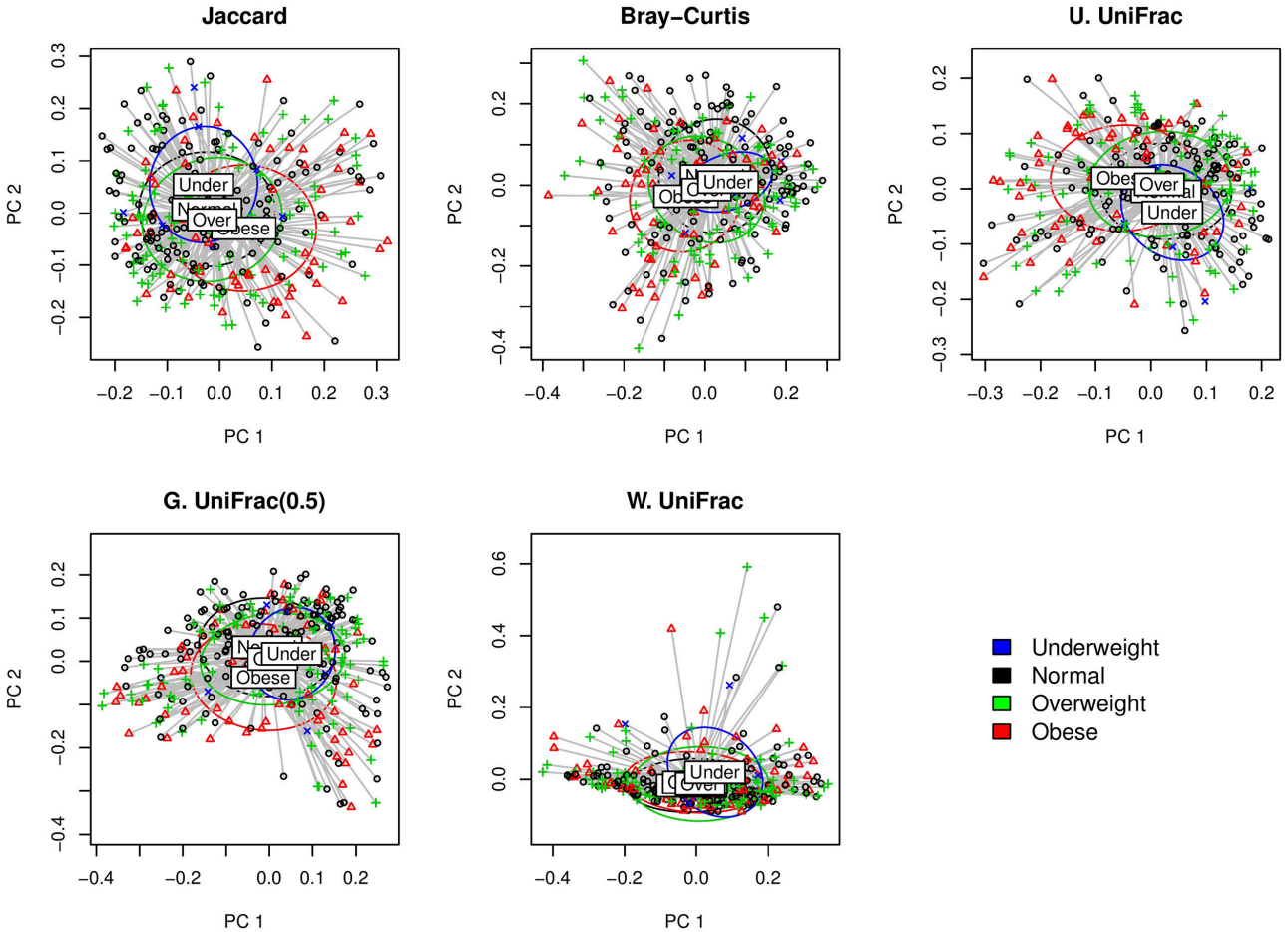


FIGURE 4 | The two-dimensional PCoA plots depicting the microbial profiles among BMI categories (i.e., Under: BMI < 18.5; Normal: 18.5 ≤ BMI < 25; Over: 25 ≤ BMI < 30; Obese: 30 ≤ BMI). Jaccard: Jaccard dissimilarity; Bray-Curtis: Bray-Curtis dissimilarity; U. UniFrac: unweighted UniFrac distance; G. UniFrac: generalized UniFrac distance ($\theta = 0.5$); W. UniFrac: weighted UniFrac distance.

0, 1, 2, and 3 course(s) of antibiotic use) adjusting for gender. We found significant association between the number of antibiotic use and microbial composition by all the item-by-item analysis [i.e., Jaccard dissimilarity (p-value: <0.001), Bray-Curtis dissimilarity (p-value: <0.001), unweighted UniFrac distance (p-value: <0.001), generalized UniFrac distance ($\theta = 0.5$) (p-value: <0.001), weighted UniFrac distance (p-value: <0.001)]. We also found the significant association for aGLMM-MiRKAT (p-value: <0.001).

DISCUSSION

In this paper, we introduced a distance-based kernel association test based on the generalized linear mixed model, GLMM-MiRKAT, for correlated (e.g., family-based or longitudinal) microbiome studies. GLMM-MiRKAT can relate microbial community composition with any type of host traits that are distributed as an exponential family distribution. Thus, GLMM-MiRKAT can be regarded as an extension of cSKAT (Zhan et al., 2018) to handle non-Gaussian host traits. Furthermore, we developed aGLMM-MiRKAT to incorporate multiple kernels for a robustly high power. aGLMM-MiKRAT is especially useful in practice, where there are various types of host traits, but our knowledge about the true association pattern is limited.

We calculate the p-values for the item-by-item GLMM-MiRKAT and aGLMM-MiRKAT using a permutation approach. The permutation approach is robust to any small or large sample size without making distributional assumptions. GLMM-MiRKAT/aGLMM-MiRKAT can be implemented for either the random intercept model or the random slope model while cSKAT is only for the random intercept model. For the random intercept model, we permute both the whole exchangeable clusters and the measurements within each cluster. We can do so because the random intercept model assumes an exchangeable (a.k.a. *compound symmetry*) within-cluster correlation structure. Therefore, for the random intercept model, our permutation approach works in any study design with either balanced or unbalanced numbers of measurements per cluster. However, for random intercept model, we permute

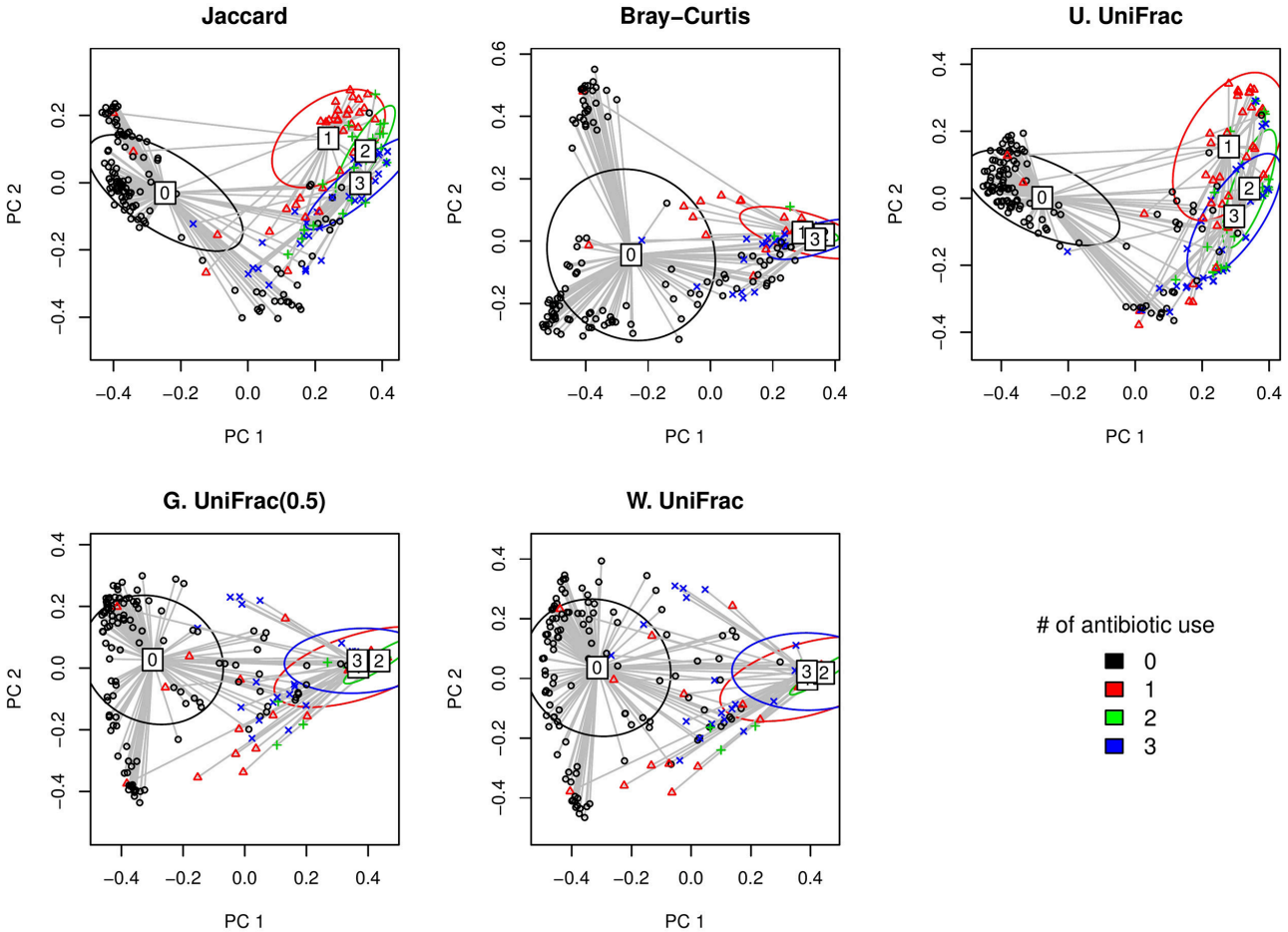


FIGURE 5 | The two-dimensional PCoA plots depicting the microbial profiles among different groups defined by the number of antibiotic use (i.e., 0, 1, 2, and 3 course(s) of antibiotic use). Jaccard: Jaccard dissimilarity; Bray-Curtis: Bray-Curtis dissimilarity; U. UniFrac: unweighted UniFrac distance; G. UniFrac: generalized UniFrac distance ($\theta = 0.5$); W. UniFrac: weighted UniFrac distance.

only the whole exchangeable clusters. Therefore, for the random slope model, our permutation approach is limited to the balanced study design with a sufficient number of whole exchangeable clusters. In practice, the random intercept model has been more widely used for many prior tests (Min and Agresti, 2005; Schifano et al., 2012; Chen et al., 2013; Zhang et al., 2014; Chen and Li, 2016; Wang et al., 2017) because the random intercepts are usually sufficient to capture the within-cluster correlation structure in responses. The model selection procedures are beyond the scope of this study and we defer the details to popular longitudinal data analysis books.

Throughout this paper, we have surveyed the bacterial kingdom as the microbial community of interest because it is usually in our shared interest (bacteria make up most of the human microbiota). However, without loss of generality, the methods can be applied to any other microbial communities, such as the kingdom of yeasts, fungi or viruses, or the lower level microbial assemblages (e.g., phyla, classes) (Koh et al., 2017). We use OTUs as the sub-units consisting of the microbial community because they are often used as the surrogate microbial species. However, any other sub-units (e.g., phylum, species, genera) can be alternatively used by researchers' choice. We considered the ecological distance measures [i.e., Jaccard dissimilarity (Jaccard, 1912), Bray-Curtis dissimilarity (Bray and Curtis, 1957) or UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007; Chen et al., 2012)] due to their popularity in the microbiome research community. However, any other distance measures or kernel matrices can be alternatively used by researcher's choice. We also make no distinction between the 16S rRNA gene sequencing (Hamady and Knight, 2009; Caporaso et al., 2010) and the shotgun metagenomic sequencing (Thomas et al., 2012) for the use of our proposed methods.

AUTHOR CONTRIBUTIONS

HK, NZ, and YL developed the method. HK performed the simulation experiments and real data analyses, and developed the software package. NZ, XZ, and JC contributed

to simulations and real data analyses. HK and NZ wrote 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 (ECHO) Data Analysis Center (U24OD023382) and Johns Hopkins University Center for AIDS Research (1P30AI094189).

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their insightful observations and comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00458/full#supplementary-material

REFERENCES

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.* 26, 32–46. doi: 10.1046/j.1442-9993.2001.01070.x

Arslan, N. (2014). Obesity, fatty liver disease and intestinal microbiota. *World J. Gastroenterol.* 20, 16452–16463. doi: 10.3748/wjg.v20.i44.16452

Bandera, A., De Benedetto, I., Bozzi, G., and Gori, A. (2018). Altered gut microbiome composition in HIV infection: causes, effects and potential intervention. *Curr. Opin. HIV AIDS* 13, 73–80. doi: 10.1097/COH.0000000000000429

Borren, N. Z., Conway, G., Garber, J. J., Khalili, H., Budree, S., Mallick, H., et al. (2018). Differences in clinical course, genetics, and the microbiome between familial and sporadic inflammatory bowel diseases. *J. Crohns. Colitis* 12, 525–531. doi: 10.1093/ecco-jcc/jjx154

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.* 27:32549. doi: 10.2307/1942268

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25. doi: 10.1080/01621459.1993.10594284

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.* 5:12. 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, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37, 196–204. doi: 10.1002/gepi.21703

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

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511801389

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, 798–799. doi: 10.1016/j.cell.2014.09.053

Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: tools, techniques. *Genome Res.* 19, 1141–1152. doi: 10.1101/gr.085464.108

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytol.* 11, 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Knights, D., Lassen, K. G., and Xavier, R. J. (2013). Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. *Gut* 62, 1505–1510. doi: 10.1136/gutjnl-2012-303954

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. doi: 10.1186/s40168-017-0262-x

Koh, H., Livanos, A. E., Blaser, M. J., and Li, H. (2018). A highly adaptive microbiome-based association test for survival traits. *BMC Genom.* 19:210. doi: 10.1186/s12864-018-4599-8

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–973. doi: 10.2307/2529876

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Lin, X. (1997). Variance component testing in generalized linear models with random effects. *Biometrika* 84, 309–326. doi: 10.1093/biomet/84.2.309

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., Koh, H., Kurtz, Z. D., Battaglia, T., PeBenito, A., Li, H., et al. (2017). Oxalobacter formigenes-associated host features and microbial community structures examined using the American Gut Project. *Microbiome* 5:108. doi: 10.1186/s40168-017-0316-0

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., 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

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

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

Min, Y., and Agresti, A. (2005). Random effect models for repeated measures for zero-inflated count data. *Stat. Model.* 5, 1–19. doi: 10.1191/1471082X05st084oa

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

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

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms*. 5, 474–504. doi: 10.1007/s10852-005-9022-1

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L., and Peyser, P. A. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* 36, 797–810. doi: 10.1002/gepi.21676

Schloss, P. D., Iverson, K. D., Petrosino, J. F., and Schloss, S. J. (2014). The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* 2:25. doi: 10.1186/2049-2618-2-25

Sneath, P. H. A., Sokal, R. R., and Freeman, W. H. (1975). Numerical taxonomy: the principles and practice of numerical classification. *Syst. Zool.* 24, 263–268. doi: 10.2307/2412767

Tang, 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

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* 2:3. doi: 10.1186/2042-5783-2-3

Wang, Z., Xu, K., Zhang, X., Wu, X., and Wang, X. (2017). Longitudinal SNP-set association analysis of quantitative phenotypes. *Genet. Epidemiol.* 41, 81–93. doi: 10.1002/gepi.22016

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., and Smith, S. M. (2015). Multi-level block permutation. *NeuroImage* 123, 253–268. doi: 10.1016/j.neuroimage.2015.05.092

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

Yang, X., Qian, Y., Xu, S., Song, Y., and Xiao, Q. (2017). Longitudinal analysis of fecal microbiome and pathologic processes in a rotenone induced mice model of Parkinson's disease. *Front. Aging Neurosci.* 9:441. doi: 10.3389/fnagi.2017.00441

Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. (2017). A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics* 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., Li, J., Krautkramer, K. A., Badri, M., Battaglia, T., Borbet, T. C., et al. (2018a). Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. *eLife* 7:e37816. doi: 10.7554/eLife.37816

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Xiangqin, C., Benson, A. K., et al. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* 18:4. doi: 10.1186/s12859-016-1441-7

Zhang, X., Pei, Y., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., et al. (2018b). Negative Binomial mixed models for analyzing longitudinal microbiome data. *Front. Microbiol.* 9:1683. doi: 10.3389/fmicb.2018.01683

Zhang, Y., Xu, Z., Shen, X., and Pan, W. (2014). Alzheimer's disease neuroimaging initiative. Testing for association with multiple traits in generalized estimating equations, with application to neuroimaging data. *NeuroImage* 96, 309–325. doi: 10.1016/j.neuroimage.2014.03.061

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

Zitvogel, L., Galluzzi, L., Viaud, S., Vétizou, M., Daillére, R., Merad, M., et al. (2015). Cancer and the gut microbiota: an unexpected link. *Sci. Transl. Med.* 7:271. doi: 10.1126/scitranslmed.3010473

Conflict of Interest Statement: 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: 17 May 2019
doi: 10.3389/fgene.2019.00454

Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals Microbiome-Dependent Relationships Between Diet and Metabolites

Zheng-Zheng Tang[1,2†], Guanhua Chen[1†], Qilin Hong[3], Shi Huang[4], Holly M. Smith[5], Rachana D. Shah[6], Matthew Scholz[7] and Jane F. Ferguson[5,8]*

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, United States, [2] Wisconsin Institute for Discovery, Madison, WI, United States, [3] Department of Statistics, University of Wisconsin–Madison, Madison, WI, United States, [4] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States, [5] Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN, United States, [6] Division of Pediatric Endocrinology, Children's Hospital of Philadelphia, Philadelphia, PA, United States, [7] Vanderbilt Technologies for Advanced Genomics (VANTAGE), Vanderbilt University Medical Center, Nashville, TN, United States, [8] Vanderbilt Translational and Clinical Cardiovascular Research Center (VTRACC), Vanderbilt University Medical Center, Nashville, TN, United States

OPEN ACCESS

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

Reviewed by:
Zhigang Li,
University of Florida, United States
Alexander Alekseyenko,
Medical University of South Carolina, United States

***Correspondence:**
Jane F. Ferguson
jane.f.ferguson@vumc.org

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *29 January 2019*
Accepted: *30 April 2019*
Published: *17 May 2019*

Citation:
Tang Z-Z, Chen G, Hong Q, Huang S, Smith HM, Shah RD, Scholz M and Ferguson JF (2019) Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals Microbiome-Dependent Relationships Between Diet and Metabolites. Front. Genet. 10:454. doi: 10.3389/fgene.2019.00454

The human microbiome has been associated with health status, and risk of disease development. While the etiology of microbiome-mediated disease remains to be fully elucidated, one mechanism may be through microbial metabolism. Metabolites produced by commensal organisms, including in response to host diet, may affect host metabolic processes, with potentially protective or pathogenic consequences. We conducted multi-omic phenotyping of healthy subjects (N = 136), in order to investigate the interaction between diet, the microbiome, and the metabolome in a cross-sectional sample. We analyzed the nutrient composition of self-reported diet (3-day food records and food frequency questionnaires). We profiled the gut and oral microbiome (16S rRNA) from stool and saliva, and applied metabolomic profiling to plasma and stool samples in a subset of individuals (N = 75). We analyzed these multi-omic data to investigate the relationship between diet, the microbiome, and the gut and circulating metabolome. On a global level, we observed significant relationships, particularly between long-term diet, the gut microbiome and the metabolome. Intake of plant-derived nutrients as well as consumption of artificial sweeteners were associated with significant differences in circulating metabolites, particularly bile acids, which were dependent on gut enterotype, indicating that microbiome composition mediates the effect of diet on host physiology. Our analysis identifies dietary compounds and phytochemicals that may modulate bacterial abundance within the gut and interact with microbiome composition to alter host metabolism.

Keywords: microbiome, diet, metabolome, multi-omics analysis, mediation, interaction

INTRODUCTION

The human microbiome is a complex ecosystem of bacteria, viruses, fungi, and bacteriophages, which interact with each other and their host (Sears, 2005; Goodman and Gordon, 2010; Minot et al., 2011). Microbiome composition is unique to an individual, is established early in life, and plays a crucial role in lifelong health (Kau et al., 2011; Minot et al., 2011; Maynard et al., 2012; Koren et al., 2013; Mohammadkhah et al., 2018). Recent discoveries implicating the microbiome in disease have been paradigm-shifting. However, we do not yet understand the molecular mechanisms linking microbiota to health status.

There is considerable site-specificity in microbiome composition, with distinct populations residing within each body site of an individual (Faust et al., 2012; Ding and Schloss, 2014). The relative contributions of the microbiota at each body site to overall host health are not yet clearly defined, but are likely be depend on both the nature of the disease, and the overall health of the host (Zhang et al., 2015). The microbiome composition of the gut is of particular interest, given its location at the crucial interface between exogenous dietary intake and internal nutrient metabolism. Translocation of microbes and microbial metabolites from the intestine to the bloodstream may occur in the absence of intestinal disease, for example during diet-induced post-prandial metabolic endotoxemia (Moreira et al., 2012; Pendyala et al., 2012; Piya et al., 2013). The gut microbiome, in combination with habitual diet, is likely to play a major role in determining gut mucosal membrane permeability and influencing systemic inflammation (Moreira et al., 2012; Pendyala et al., 2012).

Numerous factors determine the specific population of microbiota in humans, with diet being a key contributor (Zeevi et al., 2015; Ferguson et al., 2016). Specific dietary components act as substrates for microbial metabolism, shaping microbiome composition and function. Multiple macronutrient-microbiome associations have been reported, including carbohydrate intake and *Prevotella* abundance (Wu et al., 2011), saturated fat intake and *Bacteroides* and *Faecalibacterium prausnitzii*, and animal protein intake and *Bacteroides* and *Alistipes* (De Filippo et al., 2010; Cotillard et al., 2013; David et al., 2014). Microbiome composition has been linked to disease through modulation of specific metabolites and signaling pathways (Wang et al., 2011; Koeth et al., 2013; Marcobal et al., 2013; Tang et al., 2013). Gut microbial metabolism of animal-product-derived carnitine to the pro-atherogenic metabolite trimethylamine N-Oxide (TMAO) has been found to associate with increased atherosclerotic risk (Wang et al., 2011; Koeth et al., 2013). Many other dietary components may modulate disease risk through parallel mechanisms.

We hypothesized that habitual diet is associated with microbiome composition in healthy humans, and that microbiome composition is associated with gut and plasma metabolites. Using multi-omic sample analysis in up to 150 healthy subjects we profiled the microbiome (16S rRNA; stool and saliva) and the metabolome (stool and plasma) to examine the interaction between diet, the microbiome, and systemic metabolism. Our results identify global relationships and highlight novel associations between specific dietary components and circulating metabolites, that are modulated by gut bacteria, and may have consequences on health status and future disease risk.

MATERIALS AND METHODS

Study Population

The ABO Glycoproteomics in Platelets and Endothelial Cells (ABO) Study recruited healthy volunteers (N = 150; men and non-pregnant/lactating women age 18–50) to a protocol at the University of Pennsylvania from 2012–2014. Exclusion criteria included known illnesses, history of organ transplant, tobacco, and prescription medication use (except oral contraceptives). Participants were instructed to avoid over-the-counter medications, supplements, and vitamins for the 2-week period prior to the scheduled visit. Subjects provided a fasting blood sample (following a 12-h overnight fast). As part of a diet and microbiome-focused sub-study, reported here, subjects provided a stool and saliva sample for microbiome analysis (N = 136 with stool samples). All subjects completed validated 3-day food records prior to the study visit (Trabulsi and Schoeller, 2001), including on the day directly before the visit, and a weekend day. Nutrient composition was analyzed using Food Processor 8.1 (ESHA Research, Salem, OR). In addition, all subjects completed food frequency questionnaires (FFQ) to assess habitual dietary intake, including serving size, of 134 food items over the previous year [the National Cancer Institute's Diet History Questionnaire (DHQ I)] (Subar et al., 2001, 2010). Completed subject responses were analyzed using Diet*Calc version 1.5.1. Diet data were converted to nutrient intake values of 191 long-term dietary variables and 139 short-term dietary variables. All subjects provided written informed consent. The study was approved by the Institutional Review Boards of the University of Pennsylvania and Vanderbilt University.

Sample Processing, DNA Extraction and Sequencing

Subjects collected a stool sample within the 24 h prior to the study visit, using a stool collection kit (Commode Specimen Collection System, Fisher Scientific, Pittsburgh, PA, United States) provided to them. Samples were stored at 4°C and aliquots made within 36 h of sample collection. Processed samples were stored at −80°C prior to nucleic acid extraction. Subjects were instructed to brush their teeth and floss if desired, but not to use mouthwash, following their final meal on the day before the visit (>12 h before visit). Subjects were further instructed not to brush their teeth or use floss or mouthwash on the morning of their visit. Saliva samples were collected using the OMNIGene Discover OM505 DNA/RNA collection kit (DNA Genotek). Following collection, samples were divided into aliquots, and stored at −80°C prior to nucleic acid extraction. DNA was isolated from stool and saliva samples using the PSP Spin Stool DNA Plus Kit (Stratec, Germany). The 16S rRNA gene region was amplified using barcoded primers (Caporaso et al., 2012) (Eurofins Genomics, Louisville, KY, United States) and DNA libraries were cleaned

(MinElute PCR Purification kit, Qiagen, Germantown, MD, United States) prior to quantification and pooling. Pooled DNA libraries were sequenced on the MiSeq platform, 300 bp paired-end reads, at an average depth of 158,000 reads/sample (Illumina Inc., San Diego, CA, United States). Stool samples were sequenced in two batches, at the University of Pennsylvania Next-Generation Sequencing Center (UPenn NGSC, N = 107) and the Vanderbilt University Technologies for Advanced Genomics (VANTAGE) Core (N = 29). All saliva samples (N = 85) were sequenced in one batch at VANTAGE. DNA sequences in Fastq files were de-multiplexed, assembled, clustered, and phylogenetically classified using the Mothur pipeline (Schloss et al., 2009). Phylogenetic classification was performed against the Silva V123 16S database. Mothur was run using standard cutoffs, creating OTU clusters at 97% identity.

Metabolomics

Samples for a subset of individuals (N = 75 plasma and N = 75 stool, matched subjects) were profiled at Metabolon (Metabolon Inc., Morrisville, NC, United States) using their global metabolomics platform, which can identify and quantitate >1,000 metabolites through multiple mass spectrometry methods. In our study, 812 metabolites were detected in plasma, and 770 in stool samples. For each metabolite, the raw peak intensity was rescaled to set the median across all samples equal to 1, and values below the limit of detection were imputed with the lowest observed value in the dataset. Metabolite pathway enrichment analysis was conducted using MetaboAnalyst (Xia and Wishart, 2011).

Data Processing for Microbiome, Dietary and Metabolite Variables

Data processing and statistical analysis was performed in R. For the stool microbiome dataset, the OTUs were classified into 11 phyla, 20 classes, 21 orders, 32 families, and 130 genera. For the saliva microbiome dataset, the OTUs were classified into 13 phyla, 21 classes, 32 orders, 52 families, and 103 genera. We obtained two independent measures of dietary intake: 3-day food diaries (for short-term recent diet) and a food frequency questionnaire (FFQ, for long-term habitual diet). Dietary and metabolite variables were normalized using inverse normal transformation (INT) and transformed variables that did not follow a normal distribution (Shapiro–Wilk test $p < 0.05$) were removed (Maritz, 1995). These removed variables had very small variability and/or had many tied observations. The remaining dietary variables were further normalized using the residual method to adjust for total caloric intake and gender, and standardized to have mean of 0 and SD of 1. Since some dietary variables were almost identical, we chose one representative for each highly correlated cluster (Spearman correlation > 0.9), resulting in 91 long-term dietary variables and 82 short-term dietary variables in the final dataset for the downstream analysis. The complete list mapping dietary variables to the selected representative variables are available in **Supplementary Tables S1**, **S2**. In order to group metabolites that were highly correlated, we defined metabolic modules using weighted correlation network analysis WGCNA (Langfelder and Horvath, 2008). The WGCNA has been shown to be an efficient and robust method in grouping metabolomic data (McHardy et al., 2013) and allows us to summarize each module by its module eigenvalue. Using WGCNA, the gut metabolites were organized into 8 modules with 40 un-clustered metabolites, and plasma metabolites were organized into 16 modules with 169 un-clustered metabolites. The complete list of metabolites and their module organization are available in **Supplementary Tables S3**, **S4**. The abundance values of the un-clustered metabolites were combined with standardized module eigenvalues in the downstream analysis.

Distance Correlation Analysis

To evaluate the global association between pairs of high-dimensional variables among diet, microbiome and metabolomics, we used the distance correlation t-test (Székely and Rizzo, 2013) implemented in the R package "energy" to test the dependence among each pair of these three data types. Compared to Pearson correlation, the distance correlation (Székely et al., 2007; Székely and Rizzo, 2009) is a non-parametric approach (without distributional assumption) and has the power to detect general (non-linear) dependence between two sets of high- dimensional random variables. The distance correlation t-test allows the dimension of the random vectors to be larger than the sample size. The ability for detecting general dependence and handling high-dimensionality of data makes distance correlation t-test suitable for analyzing this dataset.

Microbial Enterotypes Analysis

We conducted distance-based clustering using the Partitioning Around Medoids (PAM) method (Kaufman and Rousseeuw, 1987) with the various distances including Euclidean, Bray–Curtis and Jaccard, and identified two enterotypes. To evaluate if diet-metabolite associations are modulated by microbial enterotype, we tested diet-enterotype interaction through linear regression for each pair of diet-metabolite variables, with the metabolite as the outcome, using the individual metabolites rather than metabolite modules.

Sparse Linear Log-Contrast Model

To further narrow down the interplay between diet/metabolome and microbiome, we used the sparse linear log-contrast model (Lin et al., 2014) to pinpoint important genera that are associated with dietary or metabolite variables. In this model, a dietary or metabolite variable is the response and the top 50 most abundant genera are compositional covariates. For the diet-microbiome analysis, it makes intuitive sense to analyze microbiome variables as the dependent variables since we hypothesize that diet perturbs microbial compositions. Nevertheless, we selected the log-contrast model for several reasons. It is very challenging to find a suitable probabilistic distribution for the microbial composition due to its unique features, such as zero-inflation, over-dispersion, and complex correlation structure (Li, 2015; Tang and Chen, 2018). Further, it has been demonstrated in genetic association studies that such inverse regression (treating

dependent variables as covariates) is advantageous if there are multiple dependent variables and the distribution is difficult to specify (Majumdar et al., 2016). Alternative methods that treat microbiome as dependent variables include sparse Dirichlet-Multinomial (DM) method (Chen and Li, 2013) and multivariate zero-inflated logistic-normal method (Li et al., 2018), however, we determined that the log-contrast model was the most suitable currently available model for our study. For the taxa that are unclassified at the genus level, their identities at higher levels were used. Because of the unit-sum constraint of the microbial relative abundance, the components of a composition cannot vary freely. The sparse linear log-contrast model respects the compositional nature of the microbiome data, in which the unit-sum constraint on the compositional vector is translated into the zero-sum constraint on the association coefficients across taxa in log-ratio scale (Lin et al., 2014). The zero-sum constraint is crucial for the resulting estimator to enjoy interpretive advantages over a standard lasso estimator (Tibshirani, 1994). In our analysis, we used 10-fold cross validation to choose the tuning parameter. To obtain stable selection results, we generated 100 bootstrap samples and used the same cross-validation procedure to select the genera. The genera that were selected over 70 times out of 100 were considered associated with the dietary or metabolite variable.

Microbiome Mediation Analysis

We considered how the effect of a dietary nutrient on a metabolite is transmitted through the microbial communities. Specifically, we were interested in identifying microbial taxa that mediate the diet-metabolite pathway. We focused on pairs of diet-metabolite variables linked to at least one common genus identified by the log-contrast model in section 2.7, and applied mediation analysis to the diet-gut microbiome-metabolite triplet. The top 50 most abundant genera were used as candidate microbiome mediators. To handle the compositional and high-dimensional nature of microbiome mediators, we utilized the state-of-the-art compositional mediation analysis for microbiome data (R Package ccmm) (Sohn and Li, 2019). Certain assumptions are required to make casual interpretation of the mediation effects (Imai et al., 2010; Sohn and Li, 2019). In particular, the key assumption assumes that there is no unmeasured confounding variable after controlling covariates. The method enables us to estimate the total mediation effects of microbiome composition, as well as to select important microbial taxa mediating the diet-metabolite association and estimate taxon-specific mediation effects.

RESULTS

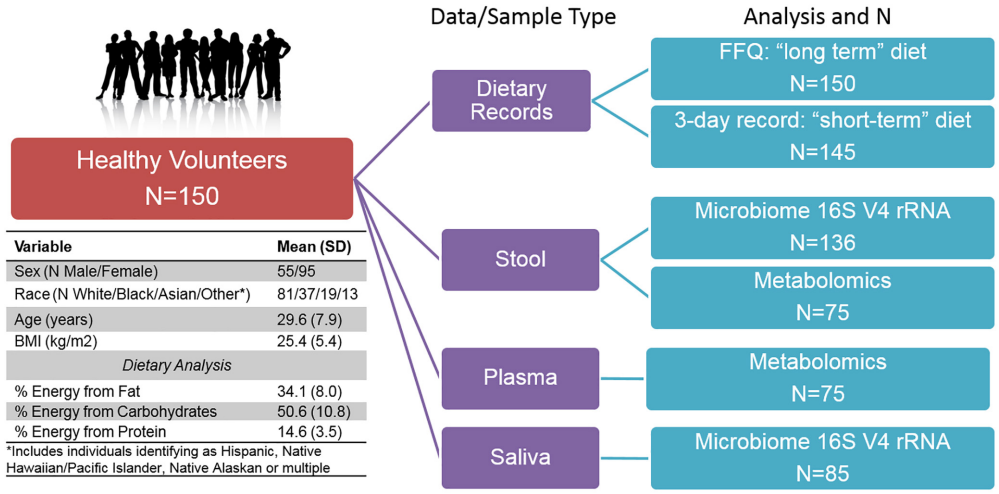
We conducted multi-omic phenotyping of up to 150 healthy subjects to probe diet, microbiome, and metabolome relationships in a cross-sectional sample. The overall study design, sample availability and subject characteristics are shown in **Figure 1**. By design, participants were healthy with no overt disease, consuming diets broadly representative of a standard American diet. Dietary variables calculated from the short and long-term diet questionnaires were significantly correlated with each other, suggesting that subjects' diets immediately prior to microbiome sampling were broadly representative of their diets over the past year. Of 150 enrolled subjects who completed a dietary questionnaire, 136 subjects provided a stool sample for microbiome analysis. We conducted metabolomic profiling in matched stool and plasma samples in a subset of these individuals (N = 75) and collected saliva samples for microbiome analysis in a separate subset (N = 85). No global associations were detected between diet, the microbiome, or metabolome, and demographic variables (age, sex, race, and BMI; PERMANOVA $p > 0.1$). We observed a difference in gut microbiome composition by batch ($p = 0.04$, UPenn vs. VANTAGE, see section "Sample Processing, DNA Extraction and Sequencing"). There were no differences in metabolite or nutrient profiles between the batches ($p > 0.1$), or in enterotype distribution (chi-square test $p = 0.86$). To assess whether the batch effect had any effect on our results, we repeated all the relevant analyses using only batch 1 samples (N = 107) and confirmed the conclusions remained the same. As the overall results did not differ, we report here the results from the analyses of the entire sample.

The Gut Microbiome Is Related to Diet and Metabolites on a Global Level

We ran a global analysis using distance correlation t-test to obtain an integrated view of the relationships and relative importance of dietary measures (short-term and long-term diet), microbiome body site samples (stool and saliva), and metabolites (stool and plasma). As shown in **Figure 2**, there were considerable inter-relationships, with particularly strong associations between the gut microbiome and the gut metabolome ($p = 2.2 \times 10^{-10}$), and between long-term diet and the gut microbiome ($p = 7.8 \times 10^{-4}$). Short-term diet was significantly associated with the gut and plasma metabolome ($p < 1 \times 10^{-3}$), but not the microbiome. We found no global associations between the saliva-derived oral microbiome and other data types. Within data types, there was very strong global correlation between short- and long-term diet ($p < 1 \times 10^{-15}$), and between stool and plasma metabolites ($p = 2.1 \times 10^{-8}$), but not between the gut and oral microbiome ($p = 0.7$). Based on the evidence in the global analysis, we decided to focus our remaining analyses on the gut microbiome and long-term diet, and to evaluate their interplay with gut and circulating metabolites.

Dietary Nutrients Are Associated With Gut Microbes

We hypothesized that gut microbiome composition would vary based on the intake of specific nutrients. From the sparse log-contrast model, we identified 61 (67%) long-term dietary nutrients associated with at least one bacterial genus (**Figure 3**). Several nutrients associated with three or more genera, as shown in **Table 1**. These dietary nutrients were



Variable	Mean (SD)
Sex (N Male/Female)	55/95
Race (N White/Black/Asian/Other*)	81/37/19/13
Age (years)	29.6 (7.9)
BMI (kg/m2)	25.4 (5.4)
Dietary Analysis	
% Energy from Fat	34.1 (8.0)
% Energy from Carbohydrates	50.6 (10.8)
% Energy from Protein	14.6 (3.5)

*Includes individuals identifying as Hispanic, Native Hawaiian/Pacific Islander, Native Alaskan or multiple

FIGURE 1 | Overview of Study design, subject characteristics, and multi-omic sample availability.

predominately found in plant-derived foods and dairy products, suggesting that inclusion or exclusion of these food groups in the diet may be particularly important in the modulation of gut microbiome composition.

Circulating and Gut Metabolites Are Associated With Gut Microbes

We hypothesized that gut microbiome composition would associate with specific metabolites in the gut and circulation, reflecting taxon-specific metabolism. We identified 123 (66%) circulating metabolite variables and modules and 34 (71%) gut metabolite variables and modules that associated with at least one bacterial genus (**Figures 4**, **5**). Several metabolites were associated with multiple genera, as shown in **Table 2**. Of these highly bacterial-related metabolites, many have known functions in bile acid metabolism, lipid and amino acid metabolism, or metabolism of xenobiotics, highlighting the important role of microbes in modulating host metabolism in key pathways.

Gut Bacterial Taxa Mediate the Association Between Dietary Nutrients and Metabolites

We were interested in whether gut bacterial taxa mediate the relationship between diet and metabolites. Mediation analysis revealed multiple taxa influencing the association between dietary intake and metabolites in plasma or stool. Given the inter-relationships between metabolic variables, we were interested in which pathways were most affected by microbiome mediation. We identified metabolic pathways with evidence for strong diet-microbiome effects, defined as having 3 or more metabolites in a sub-pathway with significant diet associations mediated by the microbiome, or association with a metabolite module (**Table 3**). These included amino acid metabolism (histidine, phenylalanine, and tyrosine), lipid metabolism (fatty acids, bile acids, and steroids), and xenobiotics (benzoate, and food components). Of the dietary variables, plant-derived nutrients (vitamins and phytochemicals) and metals were strongly represented. Our data suggest that metabolic flux through these pathways is particularly susceptible to interaction between dietary intake and microbiome composition.

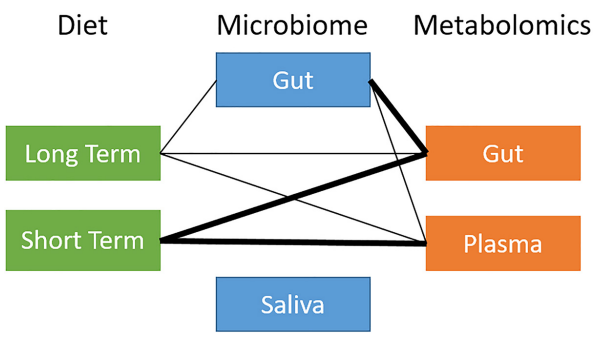


FIGURE 2 | Overview of global relationships between microbiota, diet, and metabolites. Thick line: distance correlation t-test p-value $< 10^{-5}$; thin line: distance correlation t-test $10^{-5} < p$-value $< 10^{-1}$.

Differences in Abundance of Metabolites by Gut Microbial Enterotype

We identified two gut microbiome enterotypes in our sample, with good separation of the sub-groups by Principal Coordinates Analysis (PCoA) using the Jaccard distance (see **Supplementary Figure S1**). There were 54 individuals categorized as Enterotype 1, and 82 individuals categorized as Enterotype 2. There was no difference in age or race distribution across enterotypes, or in sequencing batch, although there was a trend toward a higher proportion of women in enterotype 2 (52% vs. 69% female, chi-square test $p = 0.054$). Individuals in enterotype 2 had lower BMI (26.9 vs. 24.5, $p = 0.01$). The primary differentiating

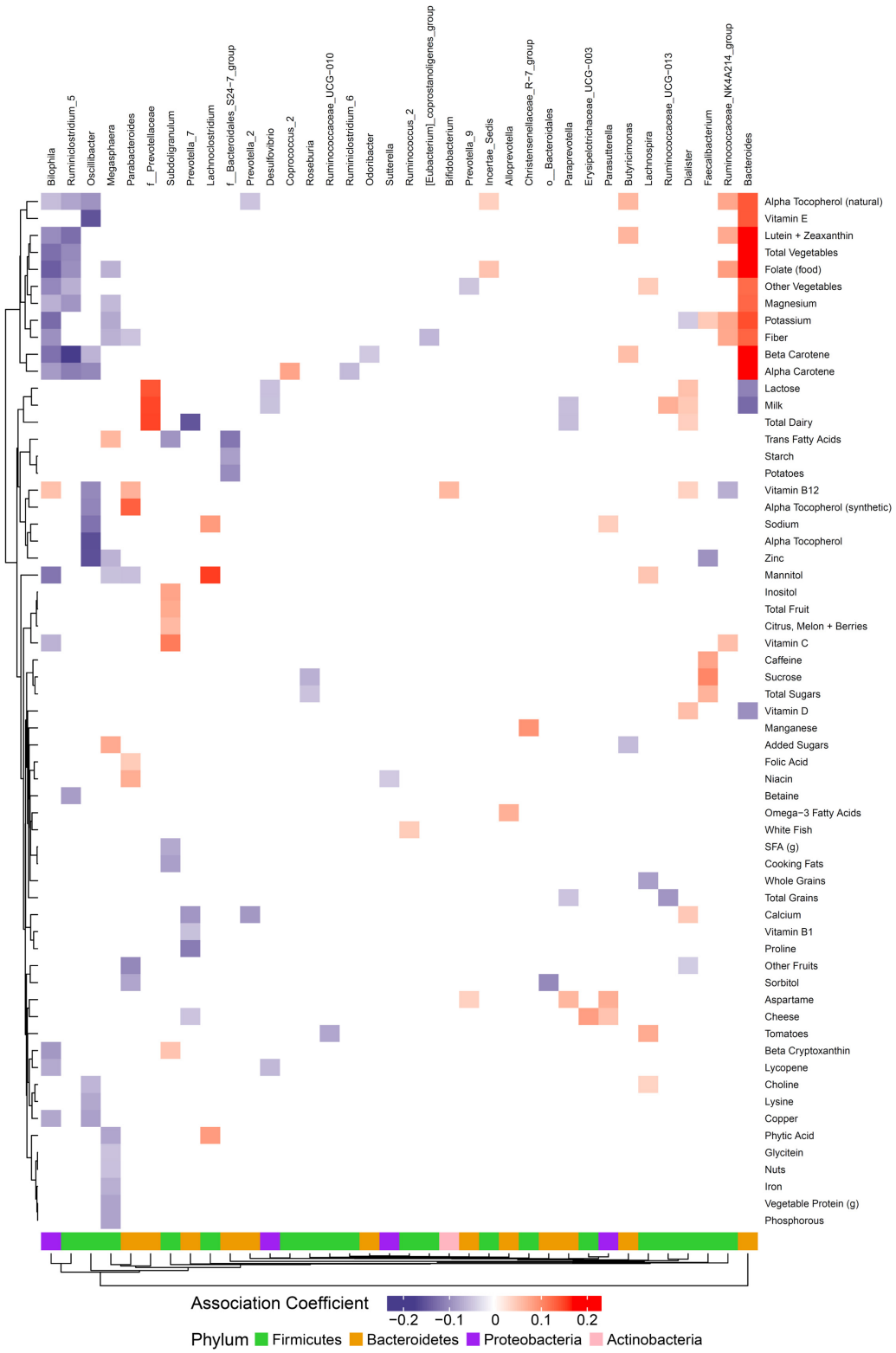


FIGURE 3 | Associations between habitual dietary nutrient intake and gut microbiome. Color intensity reflects the magnitude of the association coefficients between dietary variables and taxa.

TABLE 1 | Long term intake of dietary nutrients associated with at least three gut microbial taxa.

Dietary nutrient	Primary food source	Bacterial taxon*	
		Positive association	**Negative association**
Alpha Carotene	Plants	*Bacteroides, Coprococcus 2*	*Bilophila, Ruminiclostridium 5, Ruminiclostridium 6, Oscillibacter*
Beta Carotene		*Bacteroides, Butyricimonas*	*Bilophila, Odoribacter, Ruminiclostridium 5, Oscillibacter*
Lutein and Zeaxanthin		*Bacteroides, Ruminococcaceae NK4A214, Butyricimonas*	*Bilophila, Ruminiclostridium 5*
Vegetables		*Bacteroides, Lachnospira*	*Prevotella 9, Bilophila, Ruminiclostridium 5*
Vitamin E		*Bacteroides, IncertaeSedis, Ruminococcaceae NK4A214, Butyricimonas*	*Bilophila, Prevotella 2, Ruminiclostridium 5, Oscillibacter*
Vitamin C		*Subdoligranulum, Ruminococcaceae NK4A214*	*Bilophila*
Vitamin B12		*Parabacteroides, Bilophila, Dialister, Bifidobacterium*	*Ruminococcaceae NK4A214, Oscillibacter*
Folate		*Bacteroides, Incertae Sedis, Ruminococcaceae NK4A214*	*Bilophila, Ruminiclostridium 5, Megasphaera*
Dietary Fiber		*Bacteroides, Ruminococcaceae NK4A214*	*Parabacteroides, [Eubacterium]coprostanoligenes, Bilophila, Megasphaera*
Milk	Dairy products	*Dialister, Ruminococcaceae UCG-013,* f Prevotellaceae	*Bacteroides, Paraprevotella, Desulfovibrio*
Cheese		*Parasutterella, Erysipelotrichaceae UCG-003*	*Prevotella 7*
Calcium	Dietary Metals	*Dialister*	*Prevotella 7, Prevotella 2*
Zinc			*Faecalibacterium, Megasphaera, Oscillibacter*
Sodium		*Parasutterella, Lachnoclostridium*	*Oscillibacter*
Magnesium		*Bacteroides*	*Bilophila, Ruminiclostridium 5, Megasphaera*
Potassium		*Bacteroides, Faecalibacterium, Ruminococcaceae NK4A214*	*Bilophila, Dialister, Megasphaera*
Aspartame	Processed foods	*Prevotella 9, Parasutterella, Paraprevotella*	
Mannitol		*Lachnospira, Lachnoclostridium*	*Parabacteroides, Bilophila, Megasphaera*
Trans Fat		*Megasphaera*	*Subdoligranulum, f Bacteroidales S24-7*

**Bacterium reported as genus, unless otherwise specified (f, family; c, class; o, order; p, phylum; k, kingdom).*

characteristic between the two gut enterotypes was in the abundance of family Ruminococcaceae, with significantly higher proportion of Ruminococcaceae in enterotype 2 (**Supplementary Figure S2**). Analysis of metabolites by enterotype revealed striking differences between the groups: 112 plasma metabolites and 122 stool metabolites were significantly different by enterotype (unadjusted $p < 0.05$, **Supplementary Tables S5**, **S6**). Unadjusted p-values are reported in the enterotype analysis because the analysis used individual metabolites rather than metabolite modules and many metabolites are highly correlated. While the enterotype-associated metabolites spanned many biological pathways, they were enriched in certain categories. We selected all nominally associated metabolites for pathway enrichment analysis. Plasma metabolites that differed by enterotype were significantly enriched for amino acid metabolism ($p < 0.05$), particularly the essential amino acids phenylalanine, tryptophan, and tyrosine, the essential branched-chain amino acids valine, leucine and isoleucine, as well as arginine and proline. Stool metabolites differing by enterotype were enriched in taurine and niacin (vitamin B3) metabolism ($p < 0.05$). Individuals in Enterotype 1 had slightly higher alcohol and cholesterol consumption than Enterotype 2 ($p < 0.05$), but there were otherwise limited differences in dietary intake by enterotype, suggesting that the metabolite differences were not solely attributable to differences in diet.

Gut Microbial Enterotype Modulates the Relationship Between Diet and Metabolites

As observed in the mediation analysis for individual taxa, microbiome composition mediates the association between dietary nutrients and metabolites. We hypothesized that gut enterotype, as a composite measure of microbiome differences, would modify the relationship between dietary nutrient intake and downstream metabolism. We found evidence for significant interaction between habitual dietary intake and gut enterotype on plasma and stool metabolites across many classes of nutrients and metabolites. Of diet-metabolite pairs that were enterotype-dependent, the most frequent dietary components, which associated with >100 metabolites each, included plant-derived nutrients (fiber, carotenoids, and isoflavones) and artificial sweeteners (saccharin, mannitol, aspartame, and xylitol), as well as animal protein, trans fatty acids, caffeine, and alcohol. The diet- and enterotype-dependent metabolites spanned many

FIGURE 4 | Associations between gut microbiome and metabolites in plasma. Color intensity reflects the magnitude of the association coefficients between metabolites and taxa.

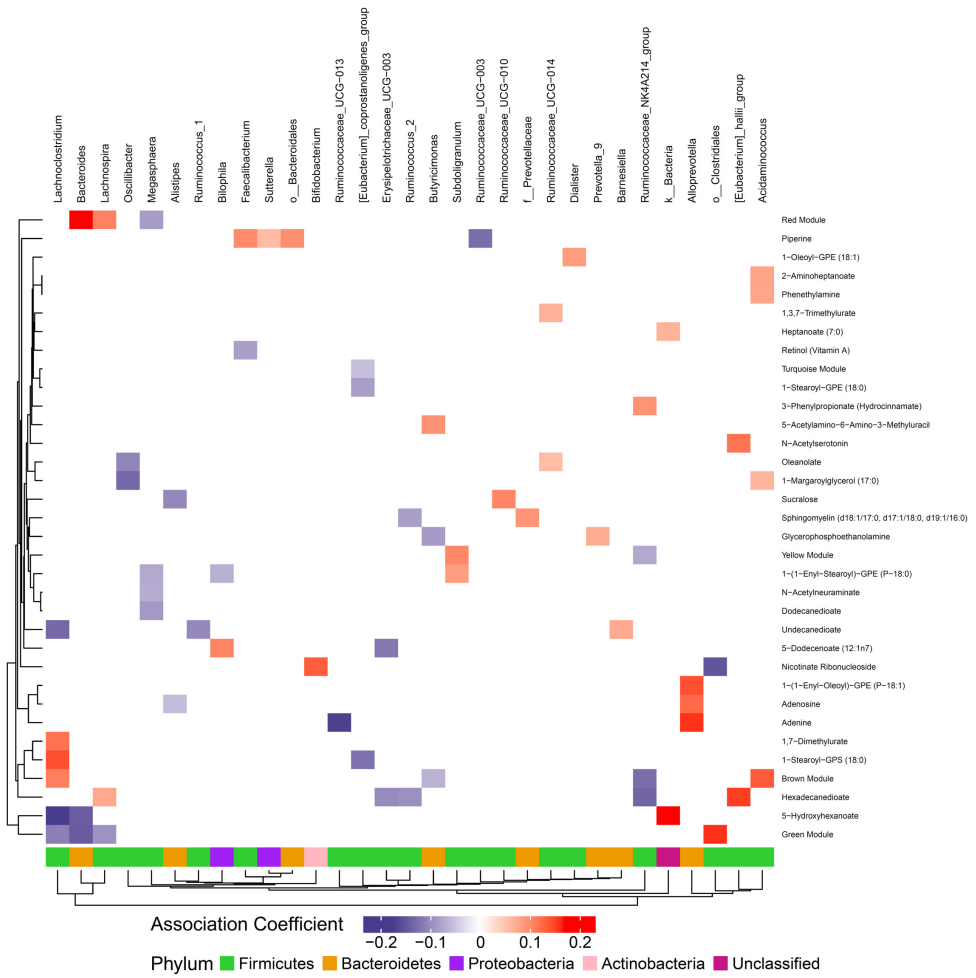


FIGURE 5 | Associations between gut microbiome and metabolites in stool. Color intensity reflects the magnitude of the association coefficients between metabolites and taxa.

pathways, but the metabolites with the most frequent associations with dietary variables (>30 dietary variables each) were predominately bile acids and xenobiotic metabolites in plasma, and xenobiotic and amino acid metabolites in stool.

Given the importance of bile acids in both gut metabolism and cardiometabolic disease risk, we were particularly interested in the observed microbiome-mediated effects of diet on bile acid signaling. As shown in **Figure 6**, habitual intake of dietary fiber was associated with higher plasma ursodeoxycholate in individuals with enterotype 1, but there was no relationship between diet and ursodeoxycholate in enterotype 2. Conversely, high dietary fiber was associated with decreased plasma taurodeoxycholate in individuals with enterotype 1, and slightly increased levels in enterotype 2. Many of the circulating bile acids were highly correlated with each other, and as such the results for taurodeoxycholate represent similar significant associations for dietary fiber with taurocholate, taurolithocholate 3 sulfate, glycolithocholate, glycolithocholate sulfate, taurochenodeoxychlate, glycodeoxycholate, glycocholate, and glycodeoxycholate sulfate, (Spearman correlation > 0.5 for metabolite pair, and $p < 0.05$ for enterotype-mediated association with diet). Of note, dietary choline was highly correlated with dietary fiber (Spearman correlation 0.7), reflecting some overlapping food sources and dietary patterns, and similar patterns of association with bile acids were also observed for choline. Interestingly, there was a modest positive relationship between plasma ursodeoxycholate ($p < 0.05$), but not plasma taurodeoxycholate, and plasma C-Reactive Protein (CRP) and BMI in individuals with enterotype 2, but not in enterotype 1 (**Figure 7**). These data suggest that individuals with enterotype 1 have bile acid metabolism that is highly diet-responsive, whereas individuals with enterotype 2 have bile acid production which is less sensitive to differences in dietary intake, but may be more likely to relate to poor metabolic health.

DISCUSSION

The gut microbiome is recognized as a key intermediate between environmental inputs and host metabolism, however, the specific

TABLE 2 | Plasma and stool metabolites associated with three or more gut microbial taxa.

Metabolite	Metabolic function	Bacterial Taxon*	
		Positive association	**Negative association**
Plasma metabolites			
Chenodeoxycholate	Primary bile acid	*Parasutterella, Incertae Sedis*	*Phascolarctobacterium, Faecalibacterium, Ruminococcus 1*
Glycolithocholate sulfate	Secondary bile acid	*Alistipes, Parasutterella, Ruminococcaceae NK4A214*	*Acidaminococcus*
7-ketolithocholate		*Parasutterella, Ruminococcaceae UCG-013*	*Alloprevotella*
Taurolithocholate 3-sulfate		*Ruminococcaceae NK4A214*	*Acidaminococcus, Ruminiclostridium 5*
Ursodeoxycholate		*Parasutterella, Dialister, Megasphaera*	
4-cholesten-3-one	Lipid	*Parabacteroides, Odoribacter*	*Parasutterella*
5alpha-pregnan-3(alpha or beta), 20beta-diol disulfate		*Parasutterella, Dialister, Faecalibacterium, [Eubacterium]hallii*	*Acidaminococcus, f Bacteroidales S24-7*
Cortisone		*Subdoligranulum, Lachnospira, [Eubacterium]hallii*	*Ruminococcus 1, Ruminiclostridium 5*
4-hydroxy-2-oxoglutaric acid		*Paraprevotella, Odoribacter, Erysipelotrichaceae UCG-003*	*Prevotella 7, Butyricimonas*
Phenol sulfate	Amino acid	*Butyricimonas, Lachnoclostridium*	*Bifidobacterium, Megasphaera, f Prevotellaceae*
Asparagine			*Parabacteroides, [Eubacterium]coprostanoligenes, Dialister*
N-acetyl-aspartyl-glutamate (NAAG)		*Phascolarctobacterium, Subdoligranulum*	*Ruminococcus 1*
3-methylglutaconate		*Paraprevotella, f Bacteroidales S24-7*	*f Lachnospiraceae*
Indolepropionate		*Sutterella*	*f Prevotellaceae, Lachnoclostridium*
Phenylacetylglutamine		*k Bacteria*	*Lachnospira, Lachnoclostridium*
N-acetylglucosaminylasparagine		*Odoribacter, Ruminococcaceae UCG-003*	*Bifidobacterium, Faecalibacterium, f Bacteroidales S24-7*
Phosphate	Energy	*k Bacteria, f Bacteroidales S24-7*	*Ruminiclostridium 5*
Fumarate		*Megasphaera, Lachnoclostridium*	*o Bacteroidales*
N6-succinyladenosine	Nucleotide	*Coprococcus 2, Oscillibacter, Lachnoclostridium*	*Bifidobacterium, f Prevotellaceae*
3-ureidopropionate		*Sutterella*	*Parabacteroides, Erysipelotrichaceae UCG-003*
4-ethylphenylsulfate	Xenobiotic	*Acidaminococcus, Ruminococcaceae NK4A214*	*Parabacteroides, Faecalibacterium, Ruminococcus 1, Lachnoclostridium*
4-allylphenol sulfate		*Paraprevotella, Sutterella*	*f Lachnospiraceae, Parabacteroides, Megasphaera*
4-hydroxychlorothalonil		*Parasutterella, f Bacteroidales S24-7*	*Prevotella 7, Megasphaera*
Retinal		*Paraprevotella, Erysipelotrichaceae UCG-003*	*o Bacteroidales, Prevotella 7*
2,3-dihydroxyisovalerate		*Oscillibacter*	*Bifidobacterium, Ruminiclostridium 5*
2-piperidinone		*Acidaminococcus*	*Ruminiclostridium 6, Ruminococcaceae NK4A214*
Gamma-CEHC		*Ruminococcaceae UCG-013*	*Incertae Sedis, Sutterella*
Salmon module		*Parasutterella, Phascolarctobacterium, Paraprevotella*	*f Bacteroidales S24-7, Coprococcus 2*
Brown module		*Bilophila, Acidaminococcus*	*f Bacteroidales S24-7*
Pink module		*Alistipes, Phascolarctobacterium, Acidaminococcus*	
Red module		*Odoribacter, Acidaminococcus*	*Prevotella 2*
Stool metabolites			
Hexadecanedioate	Lipid	*Lachnospira, [Eubacterium]hallii*	*Ruminococcus 2, Ruminococcaceae NK4A214, Erysipelotrichaceae UCG-003*
Undecanedioate		*Barnesiella*	*Ruminococcus 1, Lachnoclostridium*
3-hydroxyhexanoate		*k Bacteria*	*Bacteroides, Lachnoclostridium*
1-(1-enyl-stearoyl)-GPE (P-18:0)		*Subdoligranulum*	*Bilophila, Megasphaera*
Piperine	Xenobiotic	*o Bacteroidales, Sutterella, Faecalibacterium*	*Ruminococcaceae UCG-003*

(Continued)

TABLE 2 | Continued

Metabolite	Metabolic function	Bacterial Taxon*	
		Positive association	**Negative association**
Brown module		*Acidaminococcus, Lachnoclostridium*	*Ruminococcaceae NK4A214, Butyricimonas*
Green module		*o Clostridiales*	*Bacteroides, Lachnospira, Lachnoclostridium*
Red module		*Bacteroides, Lachnospira*	*Megasphaera*

*Bacterium reported as genus, unless otherwise specified (f, family; c, class; o, order; p, phylum; k, kingdom).

relationship between dietary nutrients, microbiome composition, and host metabolism remains poorly understood. We conducted multi-omic profiling to probe the relationship between diet, the microbiome, and metabolism in healthy adults. We identified associations between diet, the gut microbiome and the gut and plasma metabolome at a global level and identified specific microbiome-mediated associations between diet and metabolites. Our data suggest that gut microbiome composition, both at the taxon and the enterotype level, modulates how dietary nutrients are metabolized, impacting systemic host metabolism with potential downstream consequences on metabolic health.

Diet, the microbiome, and the metabolome are complex, composed of multiple inter-dependent variables, which have independent and combinatorial effects. We first examined these multi-omic datasets on a global level, to understand the inter-relationships on a broad scale. Consistent with our hypothesis, diet, the gut microbiome, and the metabolome were all related to each other. We found minimal evidence of an association between the gut and oral microbiota in the same individuals, which is consistent with previous studies, which have also reported limited overlap between different body sites (Caporaso et al., 2011; Ding and Schloss, 2014). The salivary microbiome in our sample was also not strongly related to diet, or to metabolites. This may reflect both the smaller sample size for the oral microbiome, and distal relationships between the mouth and intestinal or whole-body metabolism.

We assessed subjects' diet using two independent methods, to identify the nutrients consumed shortly before microbiome sampling, and to identify habitual long-term food consumption. There was relatively high correlation between analogous dietary variables from short and long-term estimates within subjects, suggesting that participants' diets at the time of sampling were consistent with their longer-term dietary patterns. We were interested in the relative importance of day-to-day fluctuations in dietary intake compared with longer-term patterns. We found that long-term diet as assessed by FFQ was more strongly associated with the gut microbiome than the diet consumed immediately prior to sampling (generally the 3 days prior to stool elimination). This suggests a core gut microbial population, shaped by habitual diet, that remains relatively constant despite short-term dietary fluctuations. This is supported by findings from others, who have observed relative stability in gut microbiome profiles over time, particularly in adults (Yatsunenko et al., 2012; Ding and Schloss, 2014; Dubois et al., 2017; Ruggles et al., 2018). Although large shifts in diet acutely alter microbiome composition (David et al., 2014), dietary habits over time appear to be more influential in shaping the gut microbial community. Short-term diet was more strongly associated with the gut and plasma metabolome than long-term diet, independent of the microbiome. This is consistent with a model where recently-consumed nutrients are rapidly metabolized by the host, influencing what is present in the gut and circulation at any given time. However, whether these short-term dynamic changes impact longer-term health outcomes is unknown. It is likely that repeated exposures to diet and microbiome derived metabolites over longer time frames have greater impact on lifelong health status.

Of dietary variables associated with microbiome composition and exhibiting microbiome-mediated relationships with metabolites in our sample, a large proportion are derived from plant-based foods. This is consistent with our knowledge of microbiome-mediated digestion. Plants are complex food sources, and contain many diverse nutrients, some of which are already known to interact with the microbiome. Fiber is metabolized by bacteria for production of short-chain fatty acids, which not only provide energy and selective advantages to microbes, but can affect host metabolism and immunity (Furusawa et al., 2013; Vital et al., 2014; Koh et al., 2016; Maier et al., 2017). Individuals consuming diets high in plant-derived fiber have greater microbiome diversity (Schnorr et al., 2014), while diets low in fiber lead to reduced bacterial diversity (Sonnenburg et al., 2016). Many phytochemicals are selectively metabolized by gut microbiota including isoflavones (Rowland et al., 2000; Fernandez-Raudales et al., 2012), while plants are rich sources of many vitamins, including those with known microbial interaction such as Vitamin B3/Niacin (Singh et al., 2014). Symbiotic relationships between the host and the microbiome, and optimal functioning of the holobiont, are dependent on environment, with diet being the archetypal environmental variable (Postler and Ghosh, 2017). In addition to plant foods, which have long been consumed by humans, we observed inter-relationships with artificial sweeteners, which have entered the human diet in relatively recent time. Our data do not resolve whether these have positive or negative consequences on health, but indicate that shifts toward higher consumption of processed foods and lower consumption of complex plant-based foods, common to the Western diet, have potential consequences on the gut microbiota and metabolite production.

We identified many metabolites in plasma and stool that differed by microbiome composition; indeed the majority of

TABLE 3 | Diet and microbiome mediated metabolites.

Dietary nutrient	Metabolite	Tissue	Pathway	Sub-pathway	Bacterial taxon*
Copper, lysine, vitamin E	1-methylimidazoleacetate	Plasma	Amino acid	Histidine metabolism	*Oscillibacter*
Sodium, phytic acid	Imidazole propionate	Plasma			*Ruminiclostridium_6, Oscillibacter, Lachnoclostridium*
Cheese, sodium, vitamin E	*N*-acetylhistidine	Plasma			*Parasutterella, Prevotella_2*
Sugar	3-methoxytyrosine	Plasma		Phenylalanine and tyrosine metabolism	*f__Ruminococcaceae*
Vegetables, tomato	5-bromotryptophan	Plasma			*Lachnospira*
Sugar	*N*-formylphenylalanine	Plasma			*Roseburia*
Lutein + zeaxanthin	Phenol sulfate	Plasma			*Butyricimonas*
Vegetables	Thyroxine	Plasma			*Parasutterella*
Vitamin E, B carotene, folate, lutein + zeaxanthin, copper	3-carboxy-4-methyl-5-Propyl-2-furanpropanoate	Plasma	Lipid	Fatty acid, dicarboxylate	*Butyricimonas, Oscillibacter*
Vitamin E, B carotene, lutein + zeaxanthin, grains, proline, vitamin B1, dairy, calcium, cheese	4-hydroxy-2-oxoglutaric acid	Plasma			*Butyricimonas, Erysipelotrichaceae_UCG-003, Paraprevotella, Prevotella_7*
Sugar, mannitol, fiber, folate, glycitein, iron, magnesium, potassium, phosphorous, phytic acids, nuts,	Dodecanedioate	Stool			*Megasphaera*
Vitamin E, folate, lutein + zeaxanthin, cheese, tomatoes	Hexadecanedioate	Stool			*Bifidobacterium, Butyricimonas, Ruminococcaceae_UCG-002, Erysipelotrichaceae_UCG-003*
Sodium	Undecanedioate	Stool			*Prevotella_2*
Sugar, sodium, cheese, zinc	Chenodeoxycholate	Plasma		Primary bile acid metabolism	*Parasutterella, Phascolarctobacterium, Incertae_Sedis, Faecalibacterium*
SFA (g)	Cholate	Plasma			*Phascolarctobacterium*
Cheese	Glycochenodeoxycholate Glucuronide	Plasma			*Erysipelotrichaceae_UCG-003*
Vegetables	Glycochenodeoxycholate Sulfate	Plasma			*Bacteroides*
Cheese, sodium	7-Ketolithocholate	Plasma		Secondary bile acid metabolism	*Parasutterella*
Cooking fats	Glycocholenate sulfate	Plasma			*Butyricimonas*
Potassium, folate, lutein +zeaxanthin, vitamin B12, cheese, sodium	Glycolithocholate Sulfate	Plasma			*Acidaminococcus, f__Prevotellaceae, Parasutterella, Ruminococcaceae_NK4A214_group*
B carotene, folate, vegetables	Glycoursodeoxycholate	Plasma			*Megasphaera*
Potassium, folate, lutein + zeaxanthin, vitamin B12, A carotene, B carotene	Taurolithocholate 3-sulfate	Plasma			*Acidaminococcus, Ruminiclostridium_5, Ruminococcaceae_NK4A214_group*

(Continued)

TABLE 3 | Continued

Dietary nutrient	Metabolite	Tissue	Pathway	Sub-pathway	Bacterial taxon*
Sugar, fiber, folate, glycitein, magnesium, mannitol, nuts, potassium, vegetables, cheese, calcium, sodium	Ursodeoxycholate	Plasma			*Megasphaera, Parasutterella, k__Bacteria*
Zinc, calcium, lactose, dairy, cheese, sodium, sugar, fruit, potatoes, starch, trans fat, vitamin B12, vitamin D	5α-pregnan-3(α or β),20β-diol disulfate	Plasma		Steroid/sterol	*[Eubacterium]_coprostanoligenes_group, f__Bacteroidales_S24-7_group, Erysipelotrichaceae_UCG-003*
Zinc	5α-pregnan-3β,20β-diol monosulfate	Plasma			*[Eubacterium]_coprostanoligenes_group*
Cooking fats	Cortisol	Plasma			*Butyricimonas*
Fruit, cheese, sodium, niacin	4-cholesten-3-One	Plasma			*Parabacteroides*
Phytic acid, niacin, fruit, sugar	4-ethylphenylsulfate	Plasma	Xenobiotics	Benzoate metabolism	*Lachnoclostridium, Parabacteroides, Ruminococcus_1*
Total Dairy	4-hydroxyhippurate	Plasma			*Paraprevotella*
Lutein + zeaxanthin, folate, vitamin E, B carotene, copper	Methyl-4-hydroxybenzoate sulfate	Plasma			*Bilophila, Bifidobacterium*
B carotene, folate, lutein + zeaxanthin, sodium, copper, lysine, vitamin E	2,3-dihydroxyisovalerate	Plasma		Food component/plant	*Bifidobacterium*
Folate, potassium, vitamin B12	2-piperidinone	Plasma			*Acidaminococcus, Ruminococcaceae_NK4A214_group*
Sugar, sorbitol, mannitol, glycein, iron, potassium, phosphorous, nuts, fiber, folate, magnesium, vegetables, niacin, fruit, dairy	4-allylphenol sulfate	Plasma			*Megasphaera, f__Prevotellaceae, Ruminiclostridium_6, Paraprevotella, Parabacteroides*
Lutein + zeaxanthin, vitamin B12, vitamin E, folate	Methyl indole-3-acetate	Plasma			*Ruminococcaceae_NK4A214_group, Butyricimonas*
Zinc, sucrose	Piperine	Stool			*Faecalibacterium, Parasutterella*
Proline, grains, vitamin B1, dairy, cheese, sorbitol	Retinal	Plasma			*Erysipelotrichaceae_UCG-003, Paraprevotella, o__Bacteroidales*
Sorbitol	Stachydrine	Plasma			*o__Bacteroidales*
Niacin, fruit	Thymol sulfate	Plasma			*Parabacteroides*
Sorbitol	Black, blue module	Plasma	Module		*o__Bacteroidales*
Copper, folate, lutein + zeaxanthin, B carotene, vitamin E, potatoes, starch, trans fat, vitamin B12	Brown module	Plasma			*Bilophila, Butyricimonas, f__Bacteroidales_S24-7_group*
Sugar, mannitol, fiber, phytic acid, potassium, vitamin B12, B carotene, folate	Brown module	Stool			*[Eubacterium]_coprostanoligenes_group, Acidaminococcus, Ruminococcaceae_NK4A214_group*
Calcium, fruit	Cyan module	Plasma			*Dialister, f__Ruminococcaceae*
Potatoes, starch, trans fat	Green module	Plasma			*f__Bacteroidales_S24-7_group*
Fiber, vegetables, tomatoes	Green module	Stool			*Lachnospira, Lachnoclostridium, f__Prevotellaceae*

(Continued)

TABLE 3 | Continued

Dietary nutrient	Metabolite	Tissue	Pathway	Sub-pathway	Bacterial taxon*
Cheese, sodium	Lightcyan module	Plasma			*Parasutterella*
Starch, trans fat	Purple module	Plasma			*f__Bacteroidales_S24-7_group*
Vitamin E	Red module	Plasma			*Prevotella_2*
Tomatoes, vegetables	Red module	Stool			*Christensenellaceae_R-7_group, Lachnospira*
Dairy, sodium, cheese	Salmon module	Plasma			*Paraprevotella, Parasutterella, Phascolarctobacterium*
B carotene, lutein + zeaxanthin, vitamin B12, vitamin E, lysine, sodium, copper	Tan module	Plasma			*Ruminococcaceae_NK4A214_group, Oscillibacter*
Calcium, lactose, dairy, vitamin D, vitamin B12	Turquoise module	Plasma			*Dialister*
Fiber	Turquoise module	Stool			*[Eubacterium]_coprostanoligenes_group*

**Bacterium reported as genus, unless otherwise specified (f, family; c, class; o, order; p, phylum; k, kingdom).*

metabolites appeared to be influenced by diet, the microbiome, or both. These spanned many biological pathways, but metabolites that were particularly microbiome-sensitive were pathways related to bile acid metabolism, amino acid metabolism, lipid and steroid metabolism, and metabolism of xenobiotics. While direct effects on diet or microbe-derived metabolites (e.g., xenobiotics) are to be expected, our data highlight that the microbiome also modulates key host metabolic pathways of importance not only for energy metabolism, but overall host health status including immune function. The consequences of alterations in these circulating metabolites are not fully known. Microbiome metabolites have been shown to affect inflammation and immune regulation (Levy et al., 2017; Haase et al., 2018), and we observed some association between enterotype-mediated metabolism and plasma CRP. However, further studies are needed to establish consequences of chronic alterations in metabolite signaling.

Because different bacteria can have overlapping functionality, it can be helpful to collapse the taxonomic composition into related clusters, or enterotypes, to identify individuals within subgroups of similar composition. We observed many enterotype-mediated associations, amongst them, a significant effect of gut enterotype on the relationship between dietary fiber and plasma bile acids. Bile acids are key regulators of hepatic and intestinal lipid metabolism, and have been linked to inflammation and metabolic disease (Joyce and Gahan, 2016; Chávez-Talavera et al., 2017). Microbiota contribute to bile acid metabolism, transforming host-synthesized primary bile acids to secondary bile acids, while microbiome composition may itself be shaped by bile acids (Wahlström et al., 2016; Long et al., 2017). While we present data for dietary fiber, very similar results were found for dietary choline, with fiber and choline intake strong correlated in our sample. Thus, it is not clear whether the effect is specific to fiber, choline, or another phytonutrient common to the same food source. Both fiber and choline can act as substrates or inhibitors of bile acid metabolism (Corbin and Zeisel, 2012; Dziedzic et al., 2015; Wang et al., 2017), and both have been linked to microbial metabolism (LeBlanc et al., 1998; Tang et al., 2013; Mayengbam et al., 2018; Tuncil et al., 2018), suggesting that either or both plausibly lie in a causal pathway linking diet to bile acid metabolism through microbiota.

Our study had several key strengths, but also some limitations. We recruited healthy adults, and conducted deep multi-omic phenotyping, with the goal of identifying relationships between diet, the microbiome and the metabolome independent of a disease background. While this allowed for metabolic analysis independent of disease confounding or reverse causation, it did not allow us to directly assess relationships with cardiometabolic disease. However, at least half of the participants in our study are likely to develop cardiometabolic disease in later life (Benjamin et al., 2018), suggesting that even mildly elevated risk factors may predict future disease. We measured plasma CRP, as a clinically-relevant marker of inflammation, which predicts future disease risk (Ridker, 2003), and used BMI as a proxy for obesity and future metabolic risk (Van Gaal et al., 2006). Despite our modest sample size, this is one of the

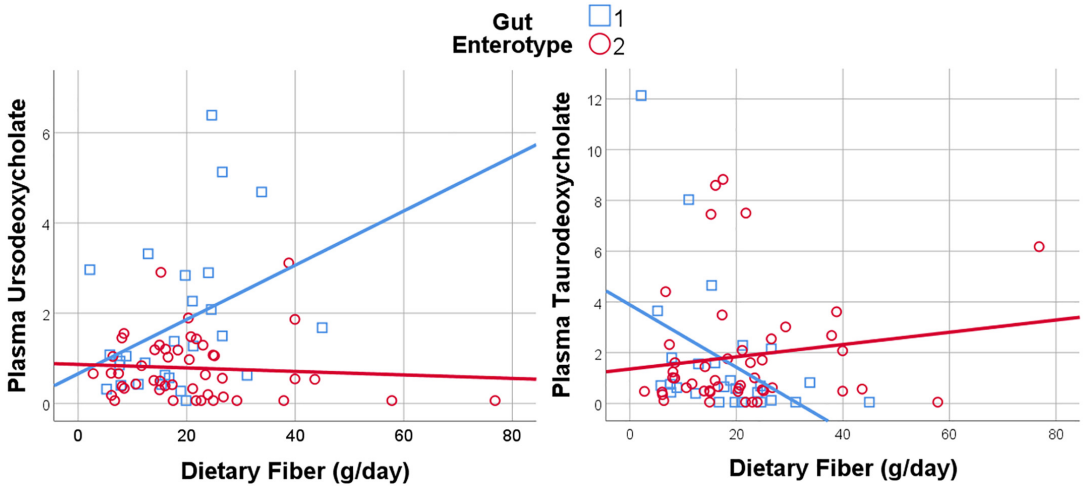


FIGURE 6 | Dietary Fiber has a gut enterotype-dependent association with plasma secondary bile acids including ursodeoxycholate and taurodeoxycholate.

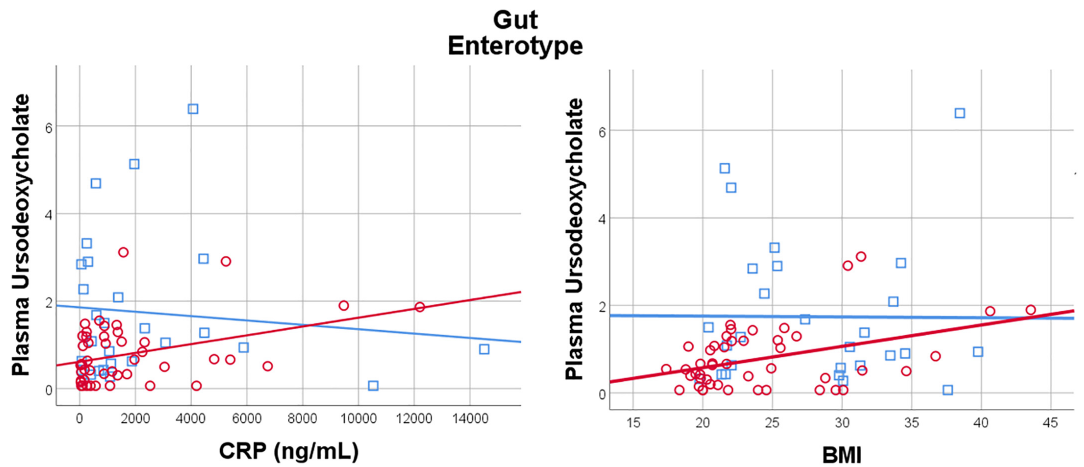


FIGURE 7 | Plasma Ursodeoxycholate has a gut enterotype-dependent relationship with plasma C-Reactive Protein and BMI, with a positive association in Enterotype 2, and no relationship in Enterotype 1.

largest studies of diet, the microbiome, and the metabolome conducted in humans. A pervasive limitation in nutritional studies is the difficulty in precise quantification of dietary intake in free-living humans. We used two independent validated dietary assessment methods, which were broadly consistent with each other, while allowing us to assess diet over different time frames. Because food is complex, and individual nutrients often co-occur in the same foods, in many cases we can not determine which food component is "causal" in a diet-microbiome-metabolite relationship. Future detailed studies to isolate individual nutrients will be required, while recognizing that nutrients exist within a complex food structure, and that an isolated nutrient (e.g., in a single supplement) may not behave the same way as a nutrient derived in conjunction with other nutrients in a food source. An important limitation of our study is the use of a single time point for data collection. While we were able to identify diet-microbiome-metabolite associations in our cross-sectional analysis, we are unable to infer causality. Future interventional studies with longitudinal sampling are required to assess relationships over time, and to determine whether changes in diet associate with microbiome-mediated changes in metabolism.

CONCLUSION

Through multi-omic analysis in a deeply-phenotyped human sample, we identified microbiome-mediated relationships between diet and circulating metabolites. Both individual microbial taxa, and microbial enterotype may relate to how dietary precursors are metabolized within the gut, and in the circulation. The potential mechanisms involved, and any long-term consequences on health status remain to be determined.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the University of Pennsylvania's clinical research standards that meet regulations relating to Good Clinical Practice (GCP). All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Institutional Review Boards of the University of Pennsylvania and Vanderbilt University.

AUTHOR CONTRIBUTIONS

JF designed the study. HS and JF performed laboratory analysis. Z-ZT, GC, QH, SH, MS, and JF performed statistical analysis. Z-ZT, GC, RS, and JF contributed to writing the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

The ABO Study was supported by U01-HL108636 and K24-HL10763 (PI: Reilly). The project was supported by an AHA Scientist Development Grant (15SDG24890015, PI: JF) and a P&F Award to JF from the Vanderbilt University Medical Center's Digestive Disease Research Center supported by NIH grant P30DK058404. The project was also supported by the Data Science Initiative Award (PI: Z-ZT) provided by the University of Wisconsin–Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00454/full#supplementary-material

REFERENCES

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., et al. (2018). Heart disease and stroke statistics-2018 update: a report from the American heart association. *Circulation* 137, e67–e492. doi: 10.1161/CIR.0000000000000558

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Chávez-Talavera, O., Tailleux, A., Lefebvre, P., and Staels, B. (2017). Bile acid control of metabolism and inflammation in obesity, type 2 diabetes, dyslipidemia, and nonalcoholic fatty liver disease. *Gastroenterology* 152, 1679–1694.e3. doi: 10.1053/j.gastro.2017.01.055

Chen, J., and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* 7, 418–442. doi: 10.1214/12-AOAS592

Corbin, K. D., and Zeisel, S. H. (2012). Choline metabolism provides novel insights into nonalcoholic fatty liver disease and its progression. *Curr. Opin. Gastroenterol.* 28, 159–165. doi: 10.1097/MOG.0b013e32834e7b4b

Cotillard, A., Kennedy, S. P., Kong, L. C., Prifti, E., Pons, N., Le Chatelier, E., et al. (2013). Dietary intervention impact on gut microbial gene richness. *Nature* 500, 585–588. doi: 10.1038/nature12480

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

Ding, T., and Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature* 509, 357–360. doi: 10.1038/nature13178

Dubois, G., Girard, C., Lapointe, F.-J., and Shapiro, B. J. (2017). The Inuit gut microbiome is dynamic over time and shaped by traditional foods. *Microbiome* 5:151. doi: 10.1186/s40168-017-0370-7

Dziedzic, K., Górecka, D., Szwengiel, A., Smoczyńska, P., Czaczyk, K., and Komolka, P. (2015). Binding of bile acids by pastry products containing bioactive substances during in vitro digestion. *Food Funct.* 6, 1011–1020. doi: 10.1039/c4fo00946k

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

Ferguson, J. F., Allayee, H., Gerszten, R. E., Ideraabdullah, F., Kris-Etherton, P. M., Ordovas, J. M., et al. (2016). Nutrigenomics, the microbiome, and gene-environment interactions: new directions in cardiovascular disease research, prevention, and treatment: a scientific statement from the American heart association. *Circ. Cardiovasc. Genet.* 9, 291–313. doi: 10.1161/HCG.0000000000000030

Fernandez-Raudales, D., Hoeflinger, J. L., Bringe, N. A., Cox, S. B., Dowd, S. E., Miller, M. J., et al. (2012). Consumption of different soymilk formulations differentially affects the gut microbiomes of overweight and obese men. *Gut Microbes* 3, 490–500. doi: 10.4161/gmic.21578

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* 504, 446–450. doi: 10.1038/nature12721

Goodman, A. L., and Gordon, J. I. (2010). Our unindicted coconspirators: human metabolism from a microbial perspective. *Cell Metab.* 12, 111–116. doi: 10.1016/j.cmet.2010.07.001

Haase, S., Haghikia, A., Wilck, N., Müller, D. N., and Linker, R. A. (2018). Impacts of microbiome metabolites on immune regulation and autoimmunity. *Immunology* 154, 230–238. doi: 10.1111/imm.12933

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* 15, 309–334. doi: 10.1037/a0020761

Joyce, S. A., and Gahan, C. G. M. (2016). Bile acid modifications at the microbe-host interface: potential for nutraceutical and pharmaceutical interventions in host health. *Annu. Rev. Food Sci. Technol.* 7, 313–333. doi: 10.1146/annurev-food-041715-033159

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Kaufman, L., and Rousseeuw, P. J. (1987). "Clustering by means of medoids," in *Statistical Data analysis Based on the L1 Norm and Related Methods*, ed. Y. Dodge (Amsterdam: Elsevier), 405–416.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* 19, 576–585. doi: 10.1038/nm.3145

Koh, A., De Vadder, F., Kovatcheva-Datchary, P., and Bäckhed, F. (2016). From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* 165, 1332–1345. doi: 10.1016/j.cell.2016.05.041

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of

microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* 9:e1002863. doi: 10.1371/journal.pcbi.1002863

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

LeBlanc, M.-J., Gavino, V., Pérea, A., Yousef, I. M., Lévy, E., and Tuchweber, B. (1998). The role of dietary choline in the beneficial effects of lecithin on the secretion of biliary lipids in rats. *Biochim. Biophys. Acta* 1393, 223–234. doi: 10.1016/S0005-2760(98)00072-1

Levy, M., Blacher, E., and Elinav, E. (2017). Microbiome, metabolites and host immunity. *Curr. Opin. Microbiol.* 35, 8–15. doi: 10.1016/j.mib.2016.10.003

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Li, Z., Lee, K., Karagas, M. R., Madan, J. C., Hoen, A. G., O'Malley, A. J., et al. (2018). Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Stat. Biosci.* 10, 587–608. doi: 10.1007/s12561-018-9219-2

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi: 10.1093/biomet/asu031

Long, S. L., Gahan, C. G. M., and Joyce, S. A. (2017). Interactions between gut bacteria and bile in health and disease. *Mol. Aspects Med.* 56, 54–65. doi: 10.1016/j.mam.2017.06.002

Maier, T. V., Lucio, M., Lee, L. H., VerBerkmoes, N. C., Brislawn, C. J., Bernhardt, J., et al. (2017). Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. *mBio* 8:e01343-17. doi: 10.1128/mBio.01343-17

Majumdar, A., Haldar, T., and Witte, J. S. (2016). Determining which phenotypes underlie a pleiotropic signal. *Genet. Epidemiol.* 40, 366–381. doi: 10.1002/gepi.21973

Marcobal, A., Kashyap, P. C., Nelson, T. A., Aronov, P. A., Donia, M. S., Spormann, A., et al. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME J.* 7, 1933–1943. doi: 10.1038/ismej.2013.89

Maritz, J. S. (1995). *Distribution-Free Statistical Methods Monographs on Statistics and Applied Probability*, 2nd Edn. Boca Raton, FL: CRC Press.

Mayengbam, S., Lambert, J. E., Parnell, J. A., Tunnicliffe, J. M., Nicolucci, A. C., Han, J., et al. (2018). Impact of dietary fiber supplementation on modulating microbiota-host-metabolic axes in obesity. *J. Nutr. Biochem.* 64, 228–236. doi: 10.1016/j.jnutbio.2018.11.003

Maynard, C. L., Elson, C. O., Hatton, R. D., and Weaver, C. T. (2012). Reciprocal interactions of the intestinal microbiota and immune system. *Nature* 489, 231–241. doi: 10.1038/nature11551

McHardy, I. H., Goudarzi, M., Tong, M., Ruegger, P. M., Schwager, E., Weger, J. R., et al. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 1:17. doi: 10.1186/2049-2618-1-17

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., et al. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 21, 1616–1625. doi: 10.1101/gr.122705.111

Mohammadkhah, A. I., Simpson, E. B., Patterson, S. G., and Ferguson, J. F. (2018). Development of the gut microbiome in children, and lifetime implications for obesity and cardiometabolic disease. *Children* 5:160. doi: 10.3390/children5120160

Moreira, A. P., Texeira, T. F., Ferreira, A. B., Peluzio Mdo, C., and Alfenas Rde, C. (2012). Influence of a high-fat diet on gut microbiota, intestinal permeability and metabolic endotoxaemia. *Br. J. Nutr.* 108, 801–809. doi: 10.1017/S0007114512001213

Pendyala, S., Walker, J. M., and Holt, P. R. (2012). A high-fat diet is associated with endotoxemia that originates from the gut. *Gastroenterology* 142, 1100–1101.e2. doi: 10.1053/j.gastro.2012.01.034

Piya, M. K., Harte, A. L., and McTernan, P. G. (2013). Metabolic endotoxaemia: is it more than just a gut feeling? *Curr. Opin. Lipidol.* 24, 78–85. doi: 10.1097/MOL.0b013e32835b4431

Postler, T. S., and Ghosh, S. (2017). Understanding the holobiont: how microbial metabolites affect human health and shape the immune system. *Cell Metab.* 26, 110–130. doi: 10.1016/j.cmet.2017.05.008

Ridker, P. M. (2003). Clinical application of C-reactive protein for cardiovascular disease detection and prevention. *Circulation* 107, 363–369. doi: 10.1161/01.cir.0000053730.47739.3c

Rowland, I. R., Wiseman, H., Sanders, T. A., Adlercreutz, H., and Bowey, E. A. (2000). Interindividual variation in metabolism of soy isoflavones and lignans: influence of habitual diet on equol production by the gut microflora. *Nutr. Cancer* 36, 27–32. doi: 10.1207/S15327914NC3601_5

Ruggles, K. V., Wang, J., Volkova, A., Contreras, M., Noya-Alarcon, O., Lander, O., et al. (2018). Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. *mSphere* 3:e00193-18. doi: 10.1128/mSphere.00193-18

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* 5:3654. doi: 10.1038/ncomms4654

Sears, C. L. (2005). A dynamic partnership: celebrating our gut flora. *Anaerobe* 11, 247–251. doi: 10.1016/j.anaerobe.2005.05.001

Singh, N., Gurav, A., Sivaprakasam, S., Brady, E., Padia, R., Shi, H., et al. (2014). Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity* 40, 128–139. doi: 10.1016/j.immuni.2013.12.007

Sohn, M., and Li, H. (2019). Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* 13, 661–681. doi: 10.1002/wsbm.1242

Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higginbottom, S. K., Wingreen, N. S., and Sonnenburg, J. L. (2016). Diet-induced extinctions in the gut microbiota compound over generations. *Nature* 529, 212–215. doi: 10.1038/nature16504

Subar, A. F., Crafts, J., Zimmerman, T. P., Wilson, M., Mittl, B., Islam, N. G., et al. (2010). Assessment of the accuracy of portion size reports using computer-based food photographs aids in the development of an automated self-administered 24-hour recall. *J. Am. Diet. Assoc.* 110, 55–64. doi: 10.1016/j.jada.2009.10.007

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., et al. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires?: the Eating at America's Table Study. *Am. J. Epidemiol.* 154, 1089–1099. doi: 10.1093/aje/154.12.1089

Székely, G. J., and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* 3, 1236–1265. doi: 10.1214/09-AOAS312

Székely, G. J., and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivar. Anal.* 117, 193–213. doi: 10.1016/j.jmva.2013.02.012

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 2769–2794. doi: 10.1214/009053607000000505

Tang, W. H., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., et al. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* 368, 1575–1584. doi: 10.1056/NEJMoa1109400

Tang, Z.-Z., and Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* doi: 10.1093/biostatistics/kxy025 [Epub ahead of print].

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Trabulsi, J., and Schoeller, D. A. (2001). Evaluation of dietary assessment instruments against doubly labeled water, a biomarker of habitual energy intake. *Am. J. Physiol.* 281, E891–E899.

Tuncil, Y. E., Thakkar, R. D., Marcia, A. D. R., Hamaker, B. R., and Lindemann, S. R. (2018). Divergent short-chain fatty acid production and succession of colonic microbiota arise in fermentation of variously-sized wheat bran fractions. *Sci. Rep.* 8:16655. doi: 10.1038/s41598-018-34912-8

Van Gaal, L. F., Mertens, I. L., and De Block, C. E. (2006). Mechanisms linking obesity with cardiovascular disease. *Nature* 444, 875–880. doi: 10.1038/nature05487

Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *mBio* 5:e00889-14. doi: 10.1128/mBio.00889-14

Wahlström, A., Sayin, S. I., Marschall, H.-U., and Bäckhed, F. (2016). Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab.* 24, 41–50. doi: 10.1016/j.cmet.2016.05.005

Wang, Y., Harding, S. V., Thandapilly, S. J., Tosh, S. M., Jones, P. J. H., and Ames, N. P. (2017). Barley β-glucan reduces blood cholesterol levels via interrupting bile acid metabolism. *Br. J. Nutr.* 118, 822–829. doi: 10.1017/S0007114517002835

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472, 57–63. doi: 10.1038/nature09922

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Xia, J., and Wishart, D. S. (2011). Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Curr. Protoc. Bioinformatics* 34, 14.10.1–14.10.48. doi: 10.1002/0471250953.bi1410s34

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094. doi: 10.1016/j.cell.2015.11.001

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* 21, 895–905. doi: 10.1038/nm.3914

Conflict of Interest Statement: 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.

The reviewer AA declared a past co-authorship with several of the authors Z-ZT and GC.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fgene.2019.00516

Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data

Grace Yoon[1], *Irina Gaynanova*[1] *and Christian L. Müller*[2]*

[1] Department of Statistics, Texas A&M University, College Station, TX, United States, [2] Center for Computational Mathematics, Flatiron Institute, New York, NY, United States

OPEN ACCESS

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

Reviewed by:
Yuan Jiang,
Oregon State University, United States
Michelle Lacey,
Tulane University, United States

***Correspondence:**
Christian L. Müller
cmueller@flatironinstitute.org

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 January 2019*
Accepted: *13 May 2019*
Published: *06 June 2019*

Citation:
Yoon G, Gaynanova I and Müller CL (2019) Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data. Front. Genet. 10:516. doi: 10.3389/fgene.2019.00516

High-throughput microbial sequencing techniques, such as targeted amplicon-based and metagenomic profiling, provide low-cost genomic survey data of microbial communities in their natural environment, ranging from marine ecosystems to host-associated habitats. While standard microbiome profiling data can provide sparse relative abundances of operational taxonomic units or genes, recent advances in experimental protocols give a more quantitative picture of microbial communities by pairing sequencing-based techniques with orthogonal measurements of microbial cell counts from the same sample. These tandem measurements provide absolute microbial count data albeit with a large excess of zeros due to limited sequencing depth. In this contribution we consider the fundamental statistical problem of estimating correlations and partial correlations from such quantitative microbiome data. To this end, we propose a semi-parametric rank-based approach to correlation estimation that can naturally deal with the excess zeros in the data. Combining this estimator with sparse graphical modeling techniques leads to the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING enables inference of statistical microbial association networks from quantitative microbiome data which can serve as high-level statistical summary of the underlying microbial ecosystem and can provide testable hypotheses for functional species-species interactions. Due to the absence of verified microbial associations we also introduce a novel quantitative microbiome data generation mechanism which mimics empirical marginal distributions of measured count data while simultaneously allowing user-specified dependencies among the variables. SPRING shows superior network recovery performance on a wide range of realistic benchmark problems with varying network topologies and is robust to misspecifications of the total cell count estimate. To highlight SPRING's broad applicability we infer taxon-taxon associations from the American Gut Project data and genus-genus associations from a

recent quantitative gut microbiome dataset. We believe that, as quantitative microbiome profiling data will become increasingly available, the semi-parametric estimators for correlation and partial correlation estimation introduced here provide an important tool for reliable statistical analysis of quantitative microbiome data.

Keywords: absolute abundance, amplicon sequencing, association network, copula, graphical model, gut microbiome, zero inflation

1. INTRODUCTION

High-throughput sequencing techniques, including targeted amplicon-based sequencing (TAS) and metagenomic profiling, provide large-scale genomic survey data of microbial communities in their natural habitats. Collaborative efforts, such as the Human Microbiome Project (HMP) (Huttenhower et al., 2012), the Earth Microbiome Project (EMP) (Bahram et al., 2018), the TARA Ocean project (Sunagawa et al., 2015), and the American Gut Project (AGP) (McDonald et al., 2018) give an increasingly detailed picture of relative abundances of operational taxonomic units, their phylogenetic relationships, and gene abundances across diverse ecosystems, ranging from marine, soil, and fresh-water to human-associated habitats albeit at different scales and resolutions. Following the seminal work in Woese and Fox (1977), TAS protocols extract and amplify specific regions in marker genes, such as the 16S rRNA gene for bacteria and archea, the 18S rRNA gene for eukaryotes, and Internal Transcribed Spacer (ITS) regions for fungi, via universal primers followed by next-generation sequencing. These profiling efforts, together with elaborate bioinformatics processing and normalization work flows (Schloss et al., 2009; Caporaso et al., 2010; Edgar, 2013; Callahan et al., 2016; Lagkouvardos et al., 2017) allow low-cost determination of highly sparse relative counts of hundreds to thousands of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) (Edgar, 2016; Callahan et al., 2017) per sample across a large number of sample sites or participants. Metagenomic profiling (Handelsman, 2004) on the other hand provide unbiased samples of the majority of genes of the sampled habitat by high-throughput shotgun sequencing. Sophisticated reference-guided as well as reference-free metagenomic read assembly, binning, and taxonomic profiling pipelines (Alneberg et al., 2014; Sczyrba et al., 2017; Sedlar et al., 2017) can, under suitable conditions on read coverage, disentangle the complex mixture of sequencing reads into entire genomes of the underlying microbes and estimate, as a high-level by-product, relative microbial abundances.

Microbiome community-level analysis tasks, such as quantifying community composition shifts across conditions or associating high-dimensional species compositions and their taxonomic profiles to each other and to environmental or host-associated covariates, require statistical estimation procedures that can handle the restrictive nature of such sparse proportional (or compositional) microbiome datasets (Li, 2015). Important examples include differential abundance techniques (McMurdie and Holmes, 2014; Mandal et al., 2015), proportionality estimation (Quinn et al., 2017), regression models with compositional covariates (Holmes et al., 2012; Lin et al., 2014), composition-adjusted correlation estimation techniques (Friedman and Alm, 2012; Cao et al., 2018), and sparse graphical models for microbial association networks (Kurtz et al., 2015; Tipton et al., 2018).

Recent advancements in microbiome profiling protocols, however, promise to alleviate the experimental shortcomings of standard TAS or metagenomic experiments by enabling a more quantitative picture of microbial communities. The experimental protocols in Gifford et al. (2011) and Satinsky et al. (2013), originally introduced for marine microbiome profiling, establish quantitative count measurements of environmental metatranscriptomic or metagenomic data by adding orthogonal internal genomic mRNA or DNA standards (of known quantity) to the environmental sample prior to sequencing. A similar spike-in approach has been proposed for gut microbiome studies in Stämmler et al. (2016). Recent quantitative approaches combine TAS techniques with robust measurements of microbial cell counts, in particular flow cytometry (Props et al., 2017; Vandeputte et al., 2017). These tandem measurements provide absolute microbial count data albeit with a large number of zero measurements due to limited sequencing depth (see **Figure 2** for an overview). Thus far, however, statistical analysis methods for these novel quantitative microbiome data remain largely elusive.

In this contribution, we consider the statistical problem of correlation and partial correlation estimation for sparse quantitative microbiome count data. To this end, we first revisit a novel semi-parametric rank-based (SPR) approach to correlation estimation that can naturally deal with the large number of zeros in the data. The SPR estimator is easy to compute and can readily replace the naïve Pearson or rank-based sample correlation estimator which are often used as a first step in downstream statistical analysis tasks, including principal component analysis, principle coordinate analysis, discriminant analysis, or canonical correlation analysis (Yoon et al., 2018). Here we use the semi-parametric rank-based estimator as a starting point for sparse partial correlation estimation and introduce the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING follows the neighborhood selection methodology outlined in Meinshausen and Bühlmann (2006) to infer the conditional dependency graph and uses stability-based model selection (Liu et al., 2010; Müller et al., 2016) to identify a sparse set of stable partial correlation estimates from quantitative microbiome data (section 2). These partial correlations can be interpreted as direct (i.e., conditionally independent) statistical microbe-microbe associations and can serve as an initial community-level description of the underlying

microbial ecosystem (Fuhrman et al., 2015; Sunagawa et al., 2015; Ruiz et al., 2017).

To evaluate our new methodology, we introduce a data generation mechanism that produces synthetic amplicon samples which exactly follow the empirical marginal cumulative distributions of measured amplicon count data while simultaneously obeying user-specified (partial) correlation dependencies among the variables and closely following user-defined total cell counts (see **Figure 2** for a summary). As ground-truth data for microbial associations remain largely elusive in current literature, our data generation mechanism might be of independent interest for testing other statistical inference schemes. We highlight SPRING's superior performance compared to standard sparse partial correlation estimation methods on a wide range of quantitative microbiome benchmark problems with varying prescribed network topologies. We also quantify, in the context of association network inference, the potential gains of quantitative over purely relative data even under misspecified totals. To showcase SPRING's broad applicability (see section 4), we first infer taxon-taxon associations from relative abundance data collected in the AGP using a pseudo-count-free log-ratio transform that can handle zero counts. Our key application is a genus-level analysis of the quantitative gut microbiome dataset put forward in Vandeputte et al. (2017). We discuss the inferred quantitative association network structure, compare it to published results, and assess, for the first time, the differences between inferred associations from measured absolute and relative abundance data in a consistent statistical framework. While we focus here on TAS-related applications, our methodology is broadly applicable to other data types with excess zeros, including quantitative metagenomics, single-cell RNA-seq, and mass spectrometry data, and thus provides a promising route toward a coherent statistical framework for correlation and partial correlation analysis of multi-omics biological data.

2. SEMI-PARAMETRIC RANK-BASED CORRELATION AND PARTIAL CORRELATION ESTIMATION

2.1. Rank-Based Estimation of Correlation Matrix for Zero-Inflated Data

A great number of multivariate statistical methods, such as principal component analysis, discriminant analysis, canonical and partial correlation analysis, to name a few, require the estimate of a covariance or correlation matrix of variables as one of the inputs. The overwhelming number of methods are based on the Pearson sample covariance matrix, which works well at capturing dependencies between variables that are normally distributed. One of the key challenges in analyzing TAS-based microbial abundance data is that it is far from normal: TAS-based measurements are inherently proportional, extremely right skewed, overdispersed, and comprise a large number of zero values. Furthermore, the zeros are not always indicative of the absence of the species, but rather a result of limited sequencing depth or primer bias. For these reasons, the sample covariance matrix is not appropriate for capturing dependencies present in microbiome data. Several methods use techniques from compositional data analysis (Aitchison, 1983), including log-ratio transforms, to adjust the data prior to any estimation, and enforce different structural constraints on the correlation or inverse correlation matrix (Friedman and Alm, 2012; Kurtz et al., 2015; Cao et al., 2018). The problem of excess zeros is typically dealt with by adding a small pseudo-count or, more recently, estimating pseudo-counts from multiple samples (Cao et al., 2017). For quantitative microbiome data, however, correlation and inverse correlation estimators are not yet available. In this work we propose to take a different approach relying on the recently proposed truncated Gaussian copula framework (Yoon et al., 2018).

First, we review the Gaussian copula model, which is sometimes referred to as non-paranormal (NPN) model (Liu et al., 2009).

Definition 1. *A random vector* $\mathbf{x} = (x_1, \ldots, x_p)^\top$ *satisfies the Gaussian copula model if there exists a set of monotonically increasing transformations* $f = (f_j)_{j=1}^p$ *satisfying* $f(\mathbf{x}) = \{f_1(x_1), \ldots, f_p(x_p)\}^\top \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ *with* $\sigma_{jj} = 1$. *We denote* $\mathbf{x} \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Sigma}, f)$.

The Gaussian copula model is commonly used in undirected graphical models (Liu et al., 2012; Fan et al., 2017) because it models the dependency between variables through the correlation matrix $\mathbf{\Sigma}$, and thus enjoys the mathematical simplicity of Gaussian multivariate distribution while relaxing the normality assumption. While the original model is only appropriate for modeling continuous variables, it has also been generalized to binary variables by adding an extra dichotomization step (Fan et al., 2017). The estimation of graphical models only requires the knowledge of the correlation matrix $\mathbf{\Sigma}$, and it has been shown (Fan et al., 2017) that consistent estimates of $\mathbf{\Sigma}$ could be easily obtained from sample Kendall's τ without the need to estimate unknown transformations f_j.

The Gaussian copula model is, however, not appropriate for quantitative microbiome data as (i) it does not take into account zero inflation, and (ii) it models continuous rather than count variables. To address (i), we take advantage of the model proposed in Yoon et al. (2018).

Definition 2 (Truncated Gaussian copula model of Yoon et al. (2018))**.** *A random vector* $\mathbf{x} = (x_1, \ldots, x_p)^\top$ *satisfies the truncated Gaussian copula model if there exists a p-dimensional random vector* $\mathbf{u} = (u_1, \ldots, u_p)^\top \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Sigma}, f)$ *such that*

$$x_j = I(u_j > c_j)u_j \quad (j = 1, \ldots, p),$$

where $I(\cdot)$ *is the indicator function and* $\mathbf{c} = (c_1, \ldots, c_p)$ *is a vector of positive constants.*

In other words, the model truncates a Gaussian copula variable so it is either zero or positive continuous. This model does not take into account that quantitative microbiome data have zeros or positive counts, but we found the continuous approximation to positive counts to work well in our simulation results (section 3).

To construct graphical models for the truncated Gaussian copula model, the estimation of the latent correlation matrix $\mathbf{\Sigma}$ is required. Yoon et al. (2018) develop a rank-based estimator for $\mathbf{\Sigma}$ by deriving the explicit form of the so-called bridge function F that connects the sample Kendall's τ estimates to the elements of $\mathbf{\Sigma}$. Given observed data $(x_{j1}, x_{k1}), \ldots, (x_{jn}, x_{kn})$ for variables j and k, the sample Kendall's τ estimate is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \text{sign}(x_{ji} - x_{ji'})\text{sign}(x_{ki} - x_{ki'}).$$

The bridge function F is defined so that $\mathbb{E}(\widehat{\tau}_{jk}) = F(\sigma_{jk})$, where σ_{jk} is the corresponding latent correlation between variables j and k. The explicit form of F for the truncated Gaussian copula model is given below.

Theorem 1 (Yoon et al. (2018)). *Let random variables x_j, x_k follow truncated Gaussian copula with corresponding latent correlation σ_{jk}. Then $\mathbb{E}(\widehat{\tau}_{jk}) = F(\sigma_{jk})$, where*

$$\begin{aligned} F(\sigma_{jk}) &= F(\sigma_{jk}; \delta_j, \delta_k) = -2\Phi_4(-\delta_j, -\delta_k, 0, 0; \mathbf{\Sigma}_{4a}) \\ &+ 2\Phi_4(-\delta_j, -\delta_k, 0, 0; \mathbf{\Sigma}_{4b}), \end{aligned}$$

$\delta_j = f_j(c_j)$, $\delta_k = f_k(c_k)$, $\Phi_4(\ldots; \mathbf{\Sigma}_4)$ is the cumulative distribution function (cdf) of the four dimensional standard normal distribution with correlation matrix $\mathbf{\Sigma}_4$,

$$\mathbf{\Sigma}_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -\sigma_{jk}/\sqrt{2} \\ 0 & 1 & -\sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -\sigma_{jk}/\sqrt{2} & 1 & -\sigma_{jk} \\ -\sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & -\sigma_{jk} & 1 \end{pmatrix}$$

and

$$\mathbf{\Sigma}_{4b} = \begin{pmatrix} 1 & \sigma_{jk} & 1/\sqrt{2} & \sigma_{jk}/\sqrt{2} \\ \sigma_{jk} & 1 & \sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \sigma_{jk}/\sqrt{2} & 1 & \sigma_{jk} \\ \sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & \sigma_{jk} & 1 \end{pmatrix}.$$

Moreover, $F(\sigma_{jk})$ is strictly increasing, so the inverse function $F^{-1}(\sigma_{jk})$ exists.

Remark 1. *To give more intuition for the form of the bridge function, we provide a brief summary of the underlying derivations here. The central part is the calculation of $\mathbb{E}\left\{\text{sign}(x_{ji} - x_{ji'})\text{sign}(x_{ki} - x_{ki'})\right\}$. Due to the effect of truncation, this calculation requires separation of events leading to zero or continuous realization of x_j before the equivalence $\text{sign}\{x_{ji} - x_{ji'}\} = \text{sign}\{f_1(x_{ji}) - f_1(x_{ji'})\}$ can be applied. This separation leads to the intersection of four events concerning normal variables (two events for continuous realization of x_j and x_k, and two events corresponding to each of the sign terms), thus explaining the appearance of the four-dimensional normal cdf in the form of the bridge function.*

Theorem 1 provides a closed-form expression of the bridge function F up to the values of thresholds δ_j, which we replace with moment-based estimators $\widehat{\delta_j}$. Let n_{0j} be the observed number of exact zeros across n realizations of variable x_j. By Definitions 1 and 2,

$$\mathbb{E}(n_{0j}/n) = P(x_j = 0) = P(u_j \le c_j) = P(f(u_j) \le \delta_j) = \Phi(\delta_j).$$

We use $\widehat{\delta_j} = \Phi^{-1}(n_{0j}/n)$ instead of δ_j and can thus calculate $\widehat{\sigma}_{jk} = F^{-1}(\widehat{\tau}_{jk})$. In practice, the inverse of the bridge function $F^{-1}(\widehat{\tau}_{jk})$ is determined numerically by finding the minimizer of the quadratic function $\{F(\sigma_{jk}) - \widehat{\tau}_{jk}\}^2$, which is unique due to the strict monotonicity of the function $F(\sigma_{jk})$.

The resulting $\widehat{\sigma}_{jk}$ are used to construct an element-wise estimator $\widehat{\Sigma}$. Since element-wise estimation does not guarantee positive semidefiniteness of $\widehat{\Sigma}$, we follow the suggestion of Fan et al. (2017) and replace $\widehat{\Sigma}$ with its projection onto the cone of positive semidefinite matrices. We use the `nearPD` function in `Matrix` R package to perform this projection. For numerical stability, we also include an additional shrinkage step of the form $\widetilde{\Sigma} = (1 - \rho)\widehat{\Sigma} + \rho I$ with $\rho = 0.01$, which guarantees strict positive definiteness of the final estimate. In simulations, we found that the method performs well across a wide range of small ρ values (see **Supplementary Material** for a sensitivity analysis of the parameter ρ). The described estimation procedure for Σ is implemented within the R package `mixedCCA` (Yoon and Gaynanova, 2018), and we refer the reader to Yoon et al. (2018) for more detailed derivations.

We refer to the proposed estimator $\widetilde{\mathbf{\Sigma}}$ of the correlation matrix $\mathbf{\Sigma}$ of truncated Gaussian copula variables as the Semi-Parametric Rank-based (SPR) correlation estimator. The SPR estimator forms the basis for the undirected graphical model framework outlined below.

2.2. Sparse Graphical Models and SPRING

We next introduce the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING relies on the estimation of an undirected graphical model from data. Undirected graphical models are typically used to represent the conditional independence relationship between the variables of random vector $\mathbf{x} \in \mathbb{R}^p$, so that

$$\text{no edge between } x_j \text{ and } x_k \iff x_j \perp x_k | \mathbf{x}_{-j,-k},$$

where $\mathbf{x}_{-j,-k}$ means all components in $\mathbf{x}$ except component j and k. If the vector $\mathbf{x}$ follows a normal distribution, then conditional independence between x_j and x_k is equivalent to zero partial correlation between variables j and k. Therefore, sparse estimates of partial correlations lead to sparse conditional independence graphs. There is a rich literature on sparse estimation of partial correlations, with perhaps the most popular methods being the neighborhood selection of Meinshausen and Bühlmann (2006) (denoted by MB from here on) and the graphical lasso (Friedman et al., 2008). While the SPR estimator of the correlation matrix proposed in section 2.1 can be used in both approaches, we found the MB method to perform better than graphical lasso in numerical simulations and therefore focus on the MB method in the remainder of the paper.

The MB method takes advantage of the connection between partial correlations and regression coefficients and performs

sparse estimation of partial correlations by regressing each of the p variables on the rest, thus finding each nodes' immediate neighbors by solving a lasso problem (Tibshirani, 1996). Given column-centered and scaled data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with columns $\mathbf{x}^j$, the MB method solves for each variable j

$$\boldsymbol{\beta}^j = \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\beta_j=0}{\operatorname{argmin}} \left\{ n^{-1}\|\mathbf{x}^j - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

Rewriting the objective function leads to

$$\begin{aligned}\boldsymbol{\beta}^j &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\beta_j=0}{\operatorname{argmin}} \left\{ \boldsymbol{\beta}^\top n^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2n^{-1}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{x}^j + \lambda\|\boldsymbol{\beta}\|_1 \right\} \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\beta_j=0}{\operatorname{argmin}} \left\{ \boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{s}^j + \lambda\|\boldsymbol{\beta}\|_1 \right\},\end{aligned}$$

where, given the centering and scaling of $\mathbf{X}$, $\mathbf{S} = n^{-1}\mathbf{X}^\top\mathbf{X}$ is the sample correlation matrix with columns $\mathbf{s}^j$. Since the standard sample correlation matrix is not suited for capturing dependencies in sparse quantitative microbiome data, SPRING replaces the sample correlation $\mathbf{S}$ in the MB method with the SPR estimator $\widetilde{\boldsymbol{\Sigma}}$ from section 2.1. The MB method comprises the regularization parameter λ which balances the trade-off between sparsity of the neighborhood and goodness of fit, and thus requires data-driven tuning. We here consider a stability-based model selection method, the Stability Approach to Regularization Selection (StARS) (Liu et al., 2010), which has been previously proven to be suitable for graphical model selection on microbiome data (Kurtz et al., 2015; Müller et al., 2016). The StARS method selects the optimal tuning parameter by repeatedly taking subsamples of the original data, estimating the graphical model for each subsample at each λ value along a prescribed regularization path, and then calculating empirical edge selection probabilities from the subsamples. The StARS edge stability criterion uses these probabilities to assess the sum of edge variabilities for each graph along the regularization path. The optimal λ is selected based on the supplied threshold t_S, with standard values being $t_S = 0.05$ and $t_S = 0.1$ (Liu et al., 2010; Kurtz et al., 2015). The threshold value represents a bound on the allowed overall edge variability over the entire graph. Lower thresholds lead to sparser, more robust graphs. Using the selected λ value, the final graphical model is refitted on the full dataset.

In summary, SPRING comprises three major components: (i) a semi-parametric rank-based correlation estimator for zero-inflated count data, (ii) the MB method to infer sparse conditional dependencies from the estimated correlation, and (iii) a stability-based approach (StARS) for sparse and robust neighborhood selection.

2.3. Extensions to Compositional Data

An important prerequisite for SPRING to be applicable to zero-inflated data is that individual count values across samples are comparable. For TAS-based microbial abundance data this condition is not satisfied because the total read count of a sample is not related to the total number of bacteria in the sample (Vandeputte et al., 2017), thus making the counts inherently proportional quantities. While this drawback is alleviated with the novel experimental techniques for quantitative microbiome data, as discussed earlier, a large number of available datasets, including the HMP and the AGP data, are only available as proportional (or compositional) data. To make SPRING amenable to statistical association inference from relative abundance data, we rely on a novel data transformation.

One of the key challenges in working with compositional data is the presence of unit-sum constraint. For correlation estimation, a common approach (see e.g., Aitchison, 1983; Kurtz et al., 2015; Cao et al., 2018) is to first apply the centered log-ratio transform (clr) to the compositional vector of each sample $\mathbf{x}_i \in \mathbb{S}^p$

$$\mathbf{z}_i = \operatorname{clr}(\mathbf{x}_i) = [\log\{x_{i1}/g(\mathbf{x}_i)\}, \log\{x_{i2}/g(\mathbf{x}_i)\}, \ldots, \log\{x_{ip}/g(\mathbf{x}_i)\}], \tag{1}$$

where $g(\mathbf{x}_i) = (\prod_{j=1}^p x_{ij})^{1/p}$ is the geometric mean of $\mathbf{x}_i$. A correlation matrix is then estimated based on the transformed $\mathbf{z}_i$, $i = 1, \ldots, n$, rather than directly on $\mathbf{x}_i$ (Aitchison, 1983). Since TAS-based microbiome profiling data have a large number of zeros, the addition of a large number of pseudo-counts is required to modify the vector of compositions to only have non-zero proportions. Adding such pseudo-counts changes the measured non-zero proportions and masks the zeros in the data, leading to zeros and non-zeros being treated equally in subsequent analysis. In addition, the choice of the actual value of the pseudo-count can influence downstream analysis results, and mere addition of extra zero components to the compositional vector would also change the transformation.

To avoid these drawbacks and to play on the strengths of SPRING in handling excess zeros, we propose a modified clr transform (mclr) that does not require the use of pseudo-counts. The key steps of the mclr transform are described below and visualized in **Figure 1**.

Contrary to recent efforts in data-driven inference of pseudo-counts (see e.g., Cao et al., 2017; de la Cruz and Kreft, 2018 and references therein), we compute the geometric mean of each sample from positive proportions only, normalize and log-transform all non-zero proportions by using that geometric mean, and apply an identical shift operation to all non-zero components in the dataset. Specifically, let $\mathbf{x}_i \in \mathbb{S}^p$ be the vector of compositions for sample i, and for simplicity of illustration, assume that the first q elements of $\mathbf{x}_i$ are zero, and the other elements are non-zero. Then we propose to apply

$$\begin{aligned}\mathbf{z}_i = \operatorname{mclr}_\varepsilon(\mathbf{x}_i) = [0, \ldots, 0, \log\{x_{i(q+1)}/\tilde{g}(\mathbf{x}_i)\} + \varepsilon, \ldots, \\ \log\{x_{ip}/\tilde{g}(\mathbf{x}_i)\} + \varepsilon],\end{aligned} \tag{2}$$

where $\tilde{g}(\mathbf{x}_i) = (\prod_{j=q+1}^p x_{ij})^{1/(p-q)}$ is the geometric mean of the non-zero elements of $\mathbf{x}_i$. When $\varepsilon = 0$, mclr_0 corresponds to clr transform applied to non-zero proportions only (**Figure 1**, middle panel). When $\varepsilon > 0$, $\operatorname{mclr}_\varepsilon$ applies a positive shift to all non-zero compositions. To make all non-zero values strictly positive, we use the data-driven shift $\varepsilon = |z_{\min}| + c$, where $z_{\min} = \min_{ij} \log\{x_{ij}/\tilde{g}(\mathbf{x}_i)\}$ and c a positive constant with the default value $c = 1$. Alternative choices are discussed in the **Supplementary Material**. The ultimate rationale for the shift is to preserve the original ordering of the entries of the compositional vector $\mathbf{x}_i$ (with zeros being the smallest) in the transformed vector $\mathbf{z}_i$. The constraint $\varepsilon > |z_{\min}|$ ensures that $z_{i(q+1)}, \ldots, z_{ip}$ are

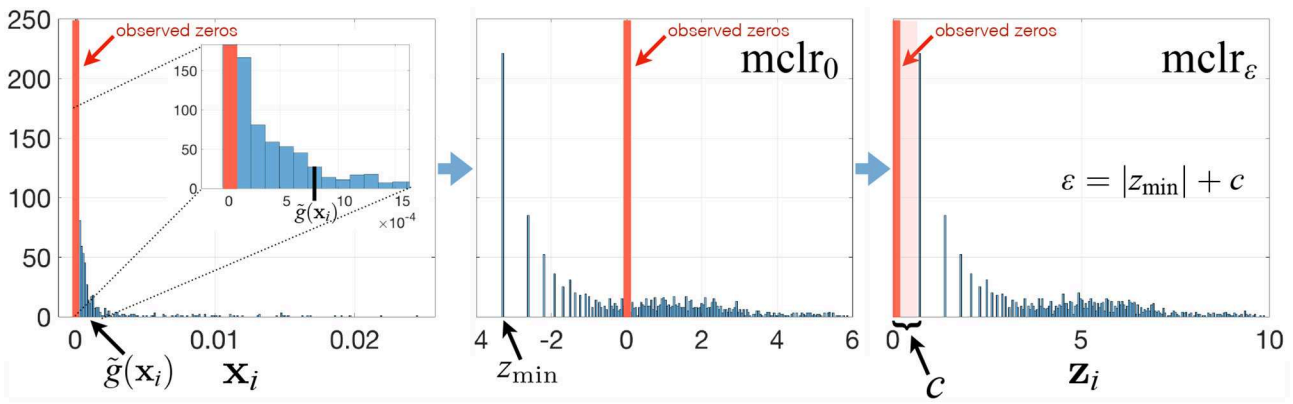


FIGURE 1 | Summary of the mclr$_\varepsilon$ transform for a relative abundance sample $\mathbf{x}_i$. See main text for a description of the different steps.

strictly positive for all i. The modified clr transform is invariant to the addition of extra zero components, preserves the original zero measurements, and is overall rank-preserving.

If a practitioner intends to infer microbial associations from relative abundance data using SPRING, we suggest to first use the mclr$_\varepsilon$ transform on relative abundance data and then apply SPRING to the transformed data. While SPRING is completely invariant to the choice of ε in mclr$_\varepsilon$ for any value of ε within the constraint due to the rank-based estimation of correlation, it does not take into the account the compositional nature of the data. Alternative ways of measuring associations between compositional components include Aitchison's variation (Aitchison, 2003), linear compositional associations (Egozcue et al., 2018), and proportionality (Quinn et al., 2017), which take the compositional constraints directly into account. Here, we will focus on correlation-based approaches and present an application of SPRING to the compositional AGP data in section 4.1.

3. SIMULATION STUDIES

3.1. Generation of Synthetic Quantitative Microbial Abundance Data

We first describe generating mechanisms for synthetic microbial abundance data with prescribed correlation or inverse correlation matrices that emulates as close as possible quantitative microbial abundance data. We closely follow ideas presented in Kurtz et al. (2015) for synthetic data generation with several important differences. The work flow of our data generation mechanism is summarized in **Figure 2**.

We propose two constructions for correlation matrices. The first construction takes directly into account the covariance of measured quantitative microbial abundance data. Given a set of n quantitative abundance samples on p taxa $\mathbf{X} \in \mathbb{R}^{n\times p}$, we compute the SPR estimator $\widetilde{\mathbf{\Sigma}}$ proposed in section 2.1 from the data and consider the resulting correlation matrix as the ground truth correlation matrix $\mathbf{\Sigma}$. The generation of synthetic samples given this correlation matrix estimate is then outlined below. Note that we do not impose any particular properties on the correlation matrix estimate, such as bounded condition number or sparsity. This construction is thus only useful for benchmarking different correlation estimation techniques.

An alternative way of generating a correlation matrix $\mathbf{\Sigma}$ is through explicitly controlling certain properties of the inverse correlation matrix. Let p be the number of nodes, i.e., the number of taxa or OTUs, and let $\mathbf{\Theta}$ be the p by p symmetric adjacency matrix such that $\theta_{ij} = 1$ if there is an edge between nodes i and j, $i \neq j$, and $\theta_{ij} = 0$ otherwise. We assume that the induced graph has no self-loops, i.e., $\theta_{ii} = 0$. We control the topology of the graph by considering three types of graph topologies: band graphs, cluster graphs, and scale-free graphs. The number of edges in the graph is denoted by e. The default value considered here is equal to twice the number of nodes ($e = 2p$), resulting in sparse graphs. Given this fixed sparsity level and the graph type, we use the R package `SpiecEasi` (Kurtz et al., 2017) to generate a precision matrix $\mathbf{\Omega}$ with the pattern of zeros corresponding to $\mathbf{\Theta}$. The non-zero entries of the lower triangular elements of $\mathbf{\Omega}$, ω_{ij} with $i > j$, are sampled uniformly at random from the intervals $[-3, -2]$ and $[2, 3]$, and the upper triangular elements are set to $\omega_{ji} = \omega_{ij}$. The diagonal elements are set to a constant such that the final precision matrix $\mathbf{\Omega}$ has a default condition number $\kappa = 100$. Using $\mathbf{\Omega}$, we generate the correlation matrix $\mathbf{\Sigma}$ by taking the inverse of the precision matrix, followed by scaling. This construction thus allows to benchmark different sparse inverse or partial correlation estimation techniques.

Given a correlation matrix $\mathbf{\Sigma}$ from either of the two constructions, we follow Kurtz et al. (2015) and use the "Normal to Anything" (NorTA) approach to generate synthetic abundance data. The NorTA method allows to generate variables with arbitrary marginal distributions from multivariate normal variables with given correlation structure. Specifically, we first generate $n \times p$ matrix $\mathbf{Z}$ with independent normal rows $\mathbf{z}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ with given correlation matrix $\mathbf{\Sigma}$, then get uniform random vectors by applying standard normal cdf transformation to each column of $\mathbf{Z}$, $\mathbf{u}^j = \Phi(\mathbf{z}^j)$ element-wise, and then apply the quantile functions of the target marginal distributions to each $\mathbf{u}^j$. In Kurtz et al. (2015), the zero-inflated negative binomial distribution (zinegbin) from `VGAM` package (Yee, 2010) is used, where the marginal distributional parameters are estimated from measured amplicon data. However, we

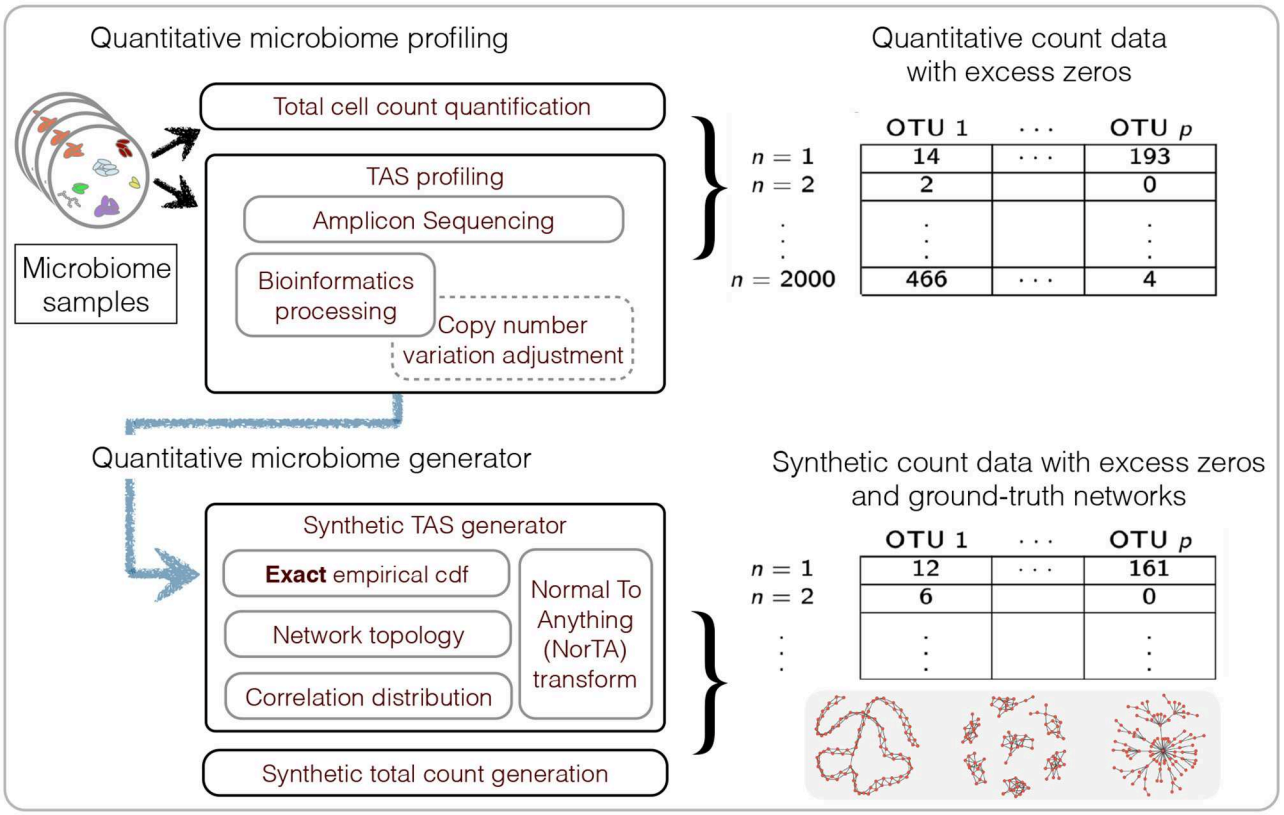


FIGURE 2 | Summary of the workflow for quantitative microbiome profiling (QMP) and the associated synthetic data generation mechanism.

found that the zinegbin distribution does not emulate well the overdispersion and skewness present in real data. This is evident by comparing the summary statistics between, e.g., the AGP data and corresponding synthetic data generated using the zinegbin, as shown in **Table 1**. To better match real amplicon data, we propose to take a different approach by using the inverse of the *empirical* cumulative distribution function (ecdf) of each OTU. This inverse can be calculated numerically by using the `uniroot.all` function in `rootSolve` package in R (Soetaert, 2009). As is evident from **Table 1**, the ecdf approach works well in mimicking the summary statistics of real TAS-based data. The match across all counts is considerably better than the match across sample abundances since the ecdf transformation is applied separately to each OTU. Although the within-sample counts are affected by the imposed correlation structure $\mathbf{\Sigma}$, the values of the sample total abundance of synthetic data with the ecdf are much closer to the measured ones than those with zinegbin. In terms of count summary statistics, the synthetic data is nearly indistinguishable from the measured data.

3.2. Estimation of Pairwise Correlations

3.2.1. Synthetic Data Generation and Methods for Comparison

We first benchmark estimation of pairwise correlations from synthetic quantitative microbial abundance data. For this purpose, we generate synthetic count data based on the quantitative microbiome profiling data, put forward in Vandeputte et al. (2017) and referred to as QMP data, and consider genus-level correlations. As the processed data used in Vandeputte et al. (2017) are not publicly available, we apply the work flow outlined in **Figure 2**. We reprocessed the available amplicon sequencing data using the standard QIIME protocol with closed-reference OTU picking (Caporaso et al., 2010), adjusted for copy number variations of the 16S rRNA gene using PICRUSt (Langille et al., 2013), filtered the data using the following three steps: (i) exclude samples whose sequencing depths (total read abundances) are $\leq$ 10000; (ii) exclude all taxa present in <30% of samples; and (iii) exclude samples whose abundance is less than the first percentile of all sequencing depths. We then combined the resulting samples with the corresponding measured total cell counts (Vandeputte et al., 2017). We next pooled $n = 106$ healthy subjects from the two available cohorts and merged all OTUs on the genus level, resulting in $p = 91$ genera. To generate synthetic data based on the QMP data with realistic correlation structure, we use the first construction method of the correlation matrix, outlined in section 3.3.1, thus considering the SPR correlation estimate on the QMP data as the ground-truth correlation matrix $\mathbf{\Sigma}$. We then generate $n = 91$ synthetic genus-level quantitative microbial abundance data that mimic the original QMP data both in terms of marginal genus distributions and correlation structure.

TABLE 1 | Comparison of summary statistics for all the counts and sample total abundance values between AGP data and two synthetic data generators.

Data	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Count data						
American Gut Project	0.0	0.0	3.0	144.9	34.0	176673.0
Synthetic (zinegbin)	0.0	0.0	33.0	170.0	125.0	54704.0
Synthetic (ecdf)	0.0	0.0	3.0	144.3	34.0	176673.0
Sample total abundance						
American Gut Project	10002.0	15149.2	20715.5	28989.0	32964.8	341632.0
Synthetic (zinegbin)	21696.0	30119.2	32543.5	33995.7	36068.0	86033.0
Synthetic (ecdf)	7354.0	18860.5	25095.5	28854.7	34285.2	196732.0

The sample size is $n = 2000$, the number of OTUs is $p = 200$, and the synthetic data is based on scale-free graph type.

In addition to the SPR correlation estimation (section 2.1) on the quantitative data, we consider three compositional correlation estimation approaches: (i) Pearson sample correlation on clr-transformed data with pseudocount addition [as used in SPIEC-EASI (Kurtz et al., 2015)], (ii) SparCC estimation from log-transformed compositions with pseudocount addition (Friedman and Alm, 2012), and (iii) SPR estimation on mclr_ε-transformed data (as described in section 2.3).

3.2.2. Results

We measure the performance of the different estimators in terms of absolute differences $|\sigma_{jk} - \widehat{\sigma}_{jk}|$, where σ_{jk} is the ground-truth correlation between genera j and k, and $\widehat{\sigma}_{jk}$ is the estimated correlation for each of the four methods. **Figure 3** shows box plots of absolute differences for the different methods. We observe that the SPR correlation estimates from the synthetic quantitative data outperform all other estimates, closely followed by the SPR estimates from mclr_ε-transformed data. SparCC and Pearson correlation on clr-transformed compositions are considerably outperformed by the SPR-type methods. The superiority of SPR-type methods is likely due to the preservation of the zero counts as zeros, thus avoiding distortions through the use of pseudo-counts, and the effective handling of the non-normality of the samples (as visible in the histogram of mclr_ε-transformed data in **Figure 1**). **Figure 4** shows the corresponding scatter plots of estimated and true pairwise correlations. We observe that SPR estimates on quantitative data are unbiased and have the smallest variance among all methods. SPR estimates on mclr_ε-transformed data have a slight downward bias and higher variance. SparCC and Pearson correlation on clr-transformed data have the worst performance both in terms of bias and variance.

3.3. Estimation of Microbial Association Networks

3.3.1. Synthetic Data Generation and Methods for Comparison

We next consider the estimation of microbial association networks. For this purpose, we generate synthetic counts from a large subset of the American Gut Project (AGP) data (McDonald et al., 2018), which comprises $p = 27116$ taxa across $n = 8440$ samples. The high dimensionality and the large sample size of the AGP data enable a more comprehensive and realistic investigation of the effects of dimensionality and sample size on the estimation of microbial associations than the QMP data. We consider the same data filtering steps as used in section 3.2.1: we (i) exclude samples whose sequencing depths (total read abundances) are ≤ 10000; (ii) exclude all taxa present in $<30\%$ of samples; and (iii) exclude samples whose abundance is less than the first percentile of all sequencing depths. This leads to a reduced dataset with $p = 481$ taxa across $n = 6482$. We consider two scenarios for the simulation studies: a large and a small sample size setting. For the large sample size setting, we randomly pick $n = 2000$ samples with total abundance at least 10,000, and then select $p = 100$ OTUs with largest abundances leading to 2000×100 matrix of synthetic counts. For the small sample size setting, we use the same strategy with $n = 500$ and $p = 200$. In the synthetic benchmarks, we treat the total observed read abundances as quantitative microbiome profiling abundances and impose sparse conditional dependencies on these counts by using the second correlation construction method, outlined in section 3.1. We refer to these samples as "True data" in the simulations. To investigate the robustness of SPRING to misspecifications of the assumed total, we also generate "Distorted data" by multiplying counts in every sample with an individual scale factor chosen uniformly at random from the interval $[0.5, 3]$. The scale factor does not affect a sample's compositional data but does distort the total abundances. The scale factor interval $[0.5, 3]$ represents a realistic distortion scenario in gut microbiome samples (see e.g., in Vandeputte et al., 2017, **Figure 2**) and is on the same order as typical fold changes of observed image-based total species counts in marine ecosystems (Ducklow, 2000). We study the performance of SPRING both on the "True" and "Distorted" synthetic data in order to assess how strongly a misspecification of the total affects association network inference.

Along with SPRING, we consider three methods for comparison. To study the influence of the sample correlation estimation, we consider the standard MB method using the Pearson sample correlation (Meinshausen and Bühlmann, 2006) [implemented in the R package huge (Zhao et al., 2012)]. We also consider two popular methods for microbial association inference from relative abundance data: SPIEC-EASI in the MB mode (Kurtz et al., 2015) and SparCC (Friedman and Alm, 2012)

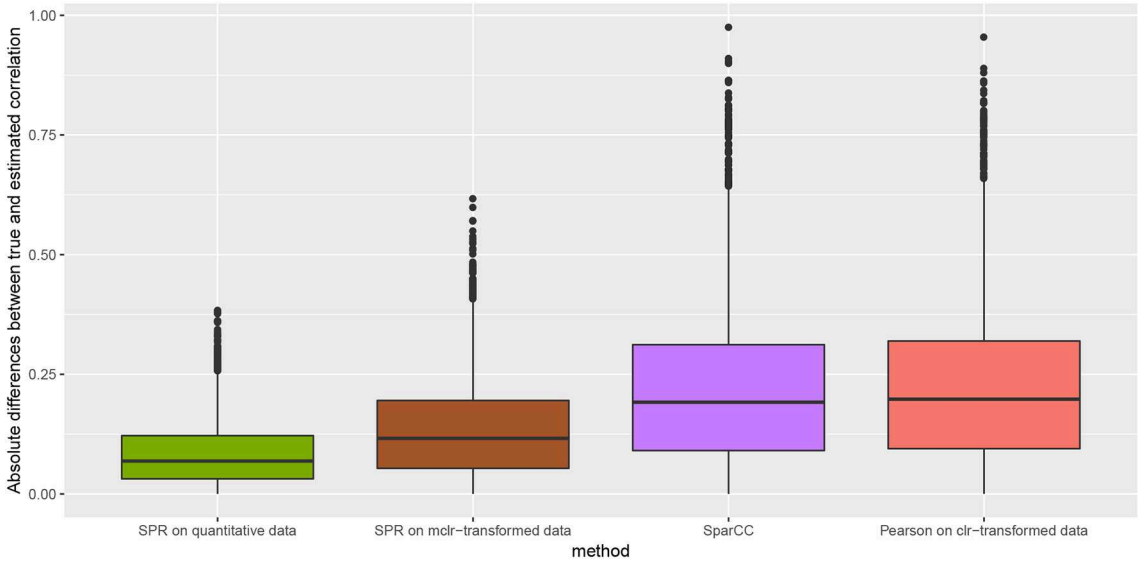


FIGURE 3 | Absolute differences between true and estimated correlation coefficients, $|\sigma_{jk} - \widehat{\sigma}_{jk}|$, for four methods: SPR correlation estimation on quantitative data (green), SPR correlation estimation on $\text{mclr}_{\varepsilon}$-transformed compositional data (brown), SparCC estimation (Friedman and Alm, 2012) (purple), and Pearson sample correlation on clr-transformed data (red) [as used in SPIEC-EASI (Kurtz et al., 2015)].

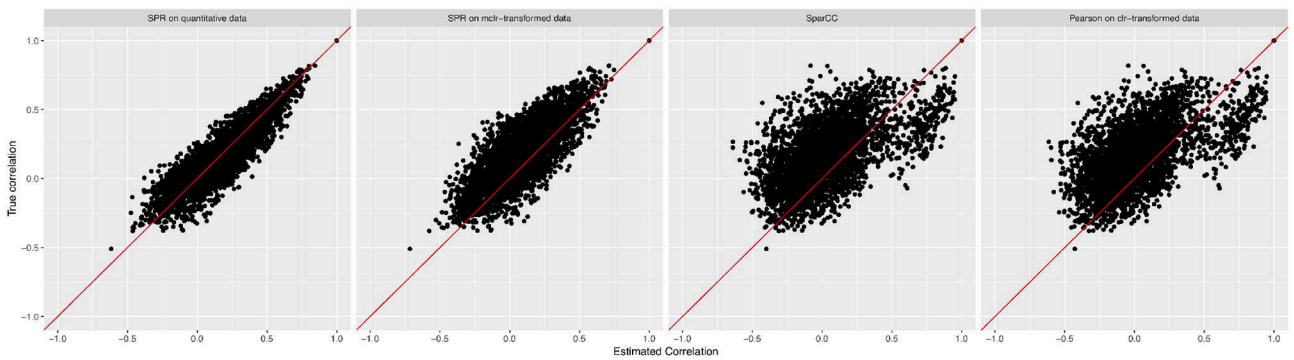


FIGURE 4 | True pairwise correlation values σ_{jk} (y-axis) vs. estimated values $\widehat{\sigma}_{jk}$ (x-axis) for four methods (from left to right): SPR correlation estimation on quantitative data, SPR correlation estimation on $\text{mclr}_{\varepsilon}$-transformed compositional data, SparCC estimation (Friedman and Alm, 2012), and Pearson sample correlation on clr-transformed data [as used in SPIEC-EASI (Kurtz et al., 2015)].

(both implemented in the R package `SpiecEasi`). The original SparCC method, however, is used for inferring marginal rather than conditional dependencies. For fair comparison with the other methods, we therefore introduce a modification of SparCC, termed invSparCC. The invSparCC method estimates the correlation matrix using the default SparCC method (as implemented in the R package `SpiecEasi`), and then uses the SparCC correlation estimator as input to the MB method, described in section 2.2. All considered methods use the neighborhood selection principle to derive a sparse graphical model, see **Table 2** for summary of all methods. The inferred adjacency and coefficient matrices are thus not guaranteed to be symmetric. We use the "or" rule and the "maxabs" rule to symmetrize the estimated adjacency and coefficient matrices, respectively. The "or" rule assigns an edge between nodes i and j if either node i is selected as a neighbor of j or node j is selected as a neighbor of i. The "maxabs" rule symmetrizes the coefficient matrix by taking the coefficient with maximum absolute value. For tuning parameter λ selection, we use the R package `pulsar` with "StARS" edge stability criterion and use 50 subsamples with subsampling ratio being fixed at $10\sqrt{n}/n$, where n is the sample size.

3.3.2. Results

We first compare the methods in terms of the Hamming distance between the true and the estimated graph. The Hamming distance is calculated as the number of edges that disagree with the true graph at each value of tuning parameter λ. The comparison of Hamming distance curves across the values of λ allows us to check the best achievable Hamming distance

value that is agnostic to tuning parameter selection scheme. We consider 50 values of λ for all methods equally spaced on a logarithmic scale, with λ_{max} corresponding to no edges in the estimated graph, and $\lambda_{min} = 0.01\lambda_{max}$. For more accurate comparison, we consider 50 replications of the data generating process for each specified combination of n and p. The mean Hamming distance values over 50 replications as functions of λ are plotted in **Figure 5**, with bands corresponding to $\pm$ two standard errors. The MB method is uniformly outperformed by all methods, confirming that standard sample correlation is not suitable for capturing dependencies in sparse quantitative microbiome data. SPIEC-EASI and invSparCC have comparable performance, with SPIEC-EASI achieving smaller mean values. SPRING performs best in all cases considered here. The most challenging scenario is the scale-free graph with low sample size, with SPRING, SPIEC-EASI, and invSparCC having comparable performance. As expected, the distortion of total abundances has no effect on the compositional methods SPIEC-EASI and invSparCC, but decreases the performance of MB and SPRING. Nevertheless, the minimum Hamming distance achieved by SPRING on distorted data is still comparable or better than the minimum distances achieved by other methods, thus suggesting that SPRING is robust to misspecification of total abundance values.

To gain further insights into the edge selection performance of the different methods, we analyze the overlapping sets of selected edges for all methods. We here focus on the cluster graph type in the low sample size regime ($n = 500$, $p = 200$). For each method we select the tuning parameter λ using StARS at $t_S = 0.1$ and repeat the experiment over 50 replications. **Figure 6** shows the average number of edges that overlap across all methods as well as average proportions of true edges among the selected ones. Among all sets uniquely identified by an individual method, SPRING shows the highest true positive rate (0.72), followed by SPIEC-EASI (0.42), invSparCC (0.12), and MB (0.01). The edge set that is jointly selected by SPRING, SPIEC-EASI, and invSparCC shows the highest true positive rate (0.95) and highest number of selected edges ($\approx$ 246), followed by the edge set jointly selected by all four methods (true positive rate 0.94 and $\approx$ 54 edges). This suggests that a promising strategy for a practitioner screening for true statistical associations is to apply SPRING, SPIEC-EASI, and invSparCC independently and select the overlapping edge set.

Next, we consider one data replication and compare the Hamming distances achieved by selecting the tuning parameter λ using StARS. The results are shown in **Figure 7** with two StARS thresholds considered (stars indicating 0.1 and circles indicating 0.05). As expected, smaller threshold corresponds to larger tuning parameter leading to sparser graph. At the same time, based on numerical results, the threshold of 0.1 tends to reach smaller Hamming distances for all methods except MB. In general, both thresholds lead to reasonable values of λ in terms of Hamming distance. As in the previous comparison, SPRING leads to smaller Hamming distance values for "True" data and is robust to misspecified total abundance values.

Finally, we compare the estimated graphs from all methods in terms of precision and recall curves, where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}};$$

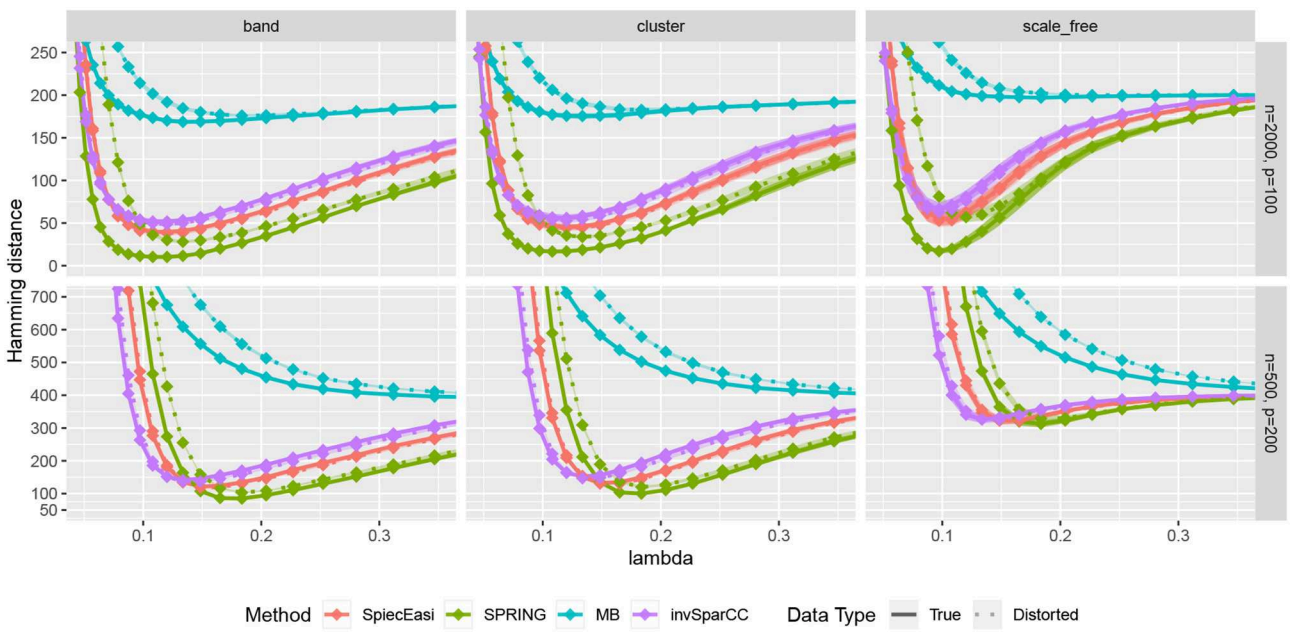


FIGURE 5 | Hamming distance as a function of tuning parameter λ. The lines correspond to mean values across 50 replications, and the bands show $\pm$ two standard errors. True abundance and distorted abundance data are distinguished by the transparency level and the line type: true data are less transparent and have solid lines; distorted data are more transparent and have dotted lines.

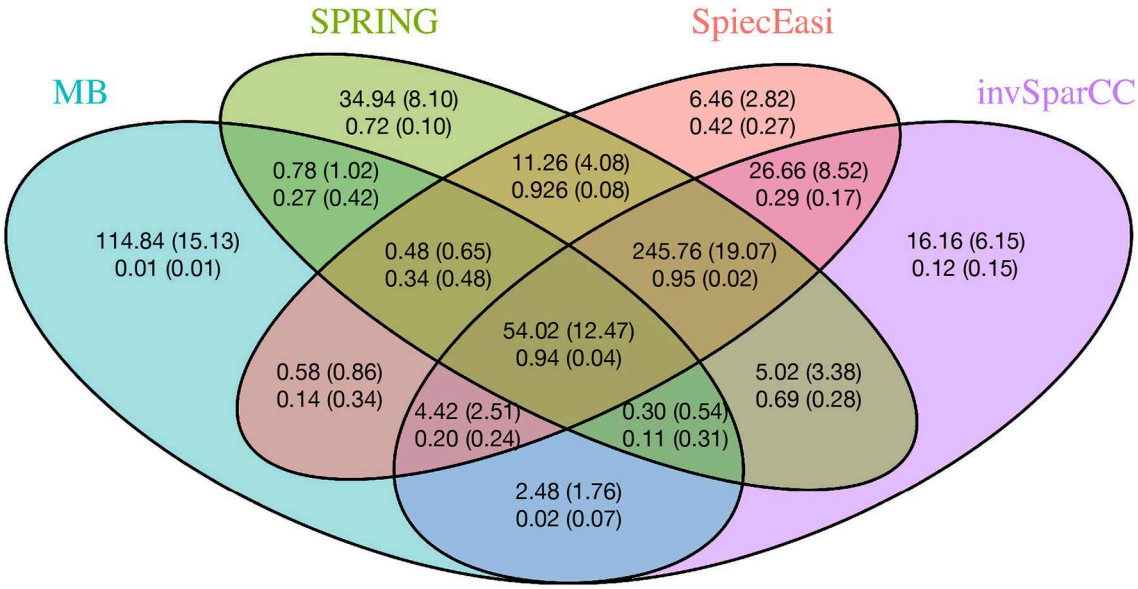


FIGURE 6 | Average number of overlapping edges (top row) and the average proportion of true edges in each corresponding overlap (bottom row) for four methods over 50 replications with $n = 500$, $p = 200$, and cluster-type graph. Corresponding standard deviations are given in parentheses.

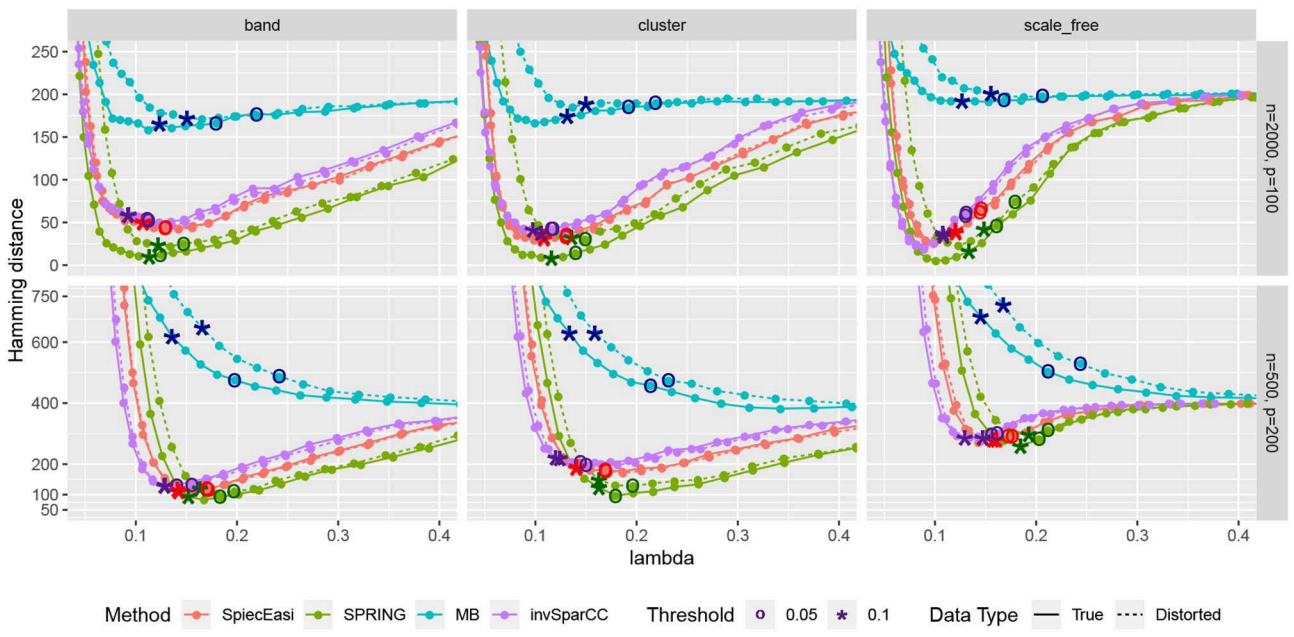


FIGURE 7 | Hamming distance as a function of tuning parameter λ. The distances at the tuning parameters selected by StARS are marked with star-shaped points ($t_S = 0.1$) and circle-shaped points ($t_S = 0.05$). True data are plotted with solid lines and distorted data are plotted with dotted lines.

TP, FP, and FN indicate the number of True Positives, False Positives, and False Negatives, respectively. To construct the curves, we extract the edge selection probabilities based on 50 subsamples from `pulsar` corresponding to tuning parameter with $t_S = 0.1$. We calculate precision and recall values by changing the threshold for edge selection probability from 1 to 0, interpolating the precision-recall values at the edges for no selection (recall = 0, precision = 1) and complete selection [recall = 1, precision = $4/(p-1)$]. Here $4/(p-1)$ is the probability of choosing true edges ($e = 2p$) at random among all possible edges ($p(p-1)/2$). The resulting curves are shown in **Figure 8**. For True data, SPRING achieves the highest precision-recall curves across all scenarios. The Area Under the Precision-Recall curve (AUPR) values are reported in **Table 3**. For the distorted data,

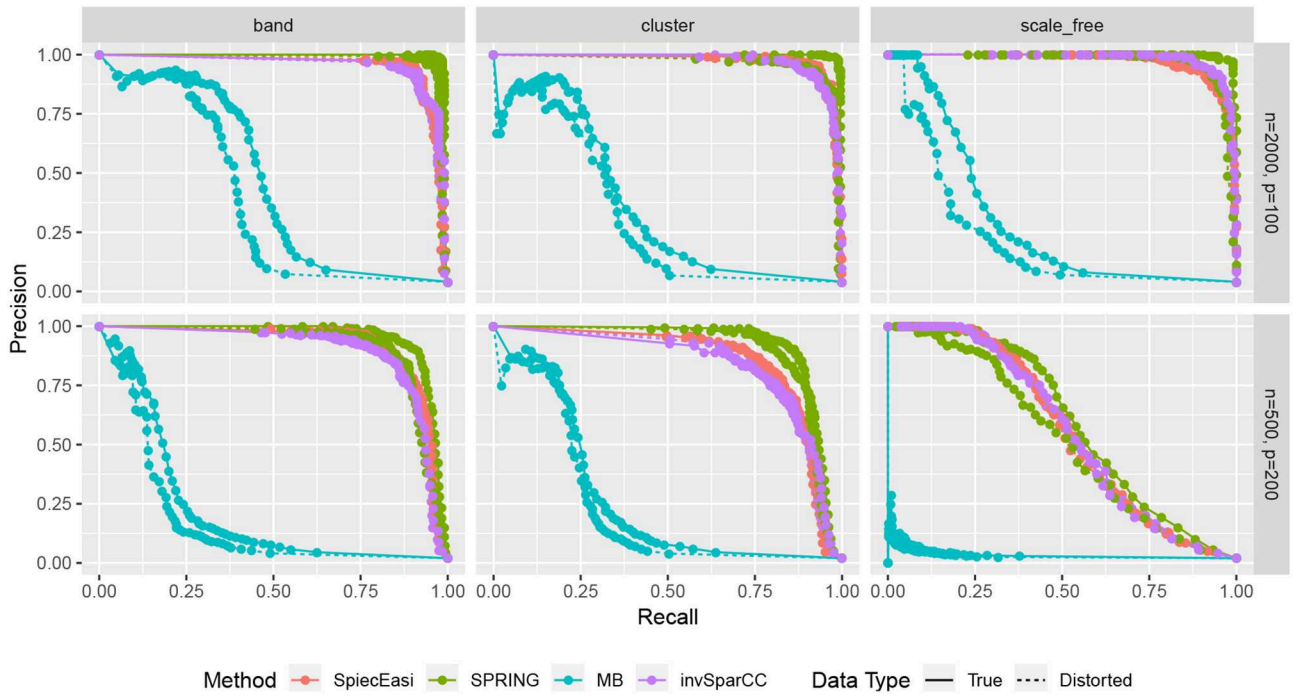


FIGURE 8 | Precision-Recall curves based on edge selection probabilities from pulsar with $t_S = 0.1$. True data are plotted with solid lines and distorted data are plotted with dotted lines.

SPRING is still best or among the best methods for band and cluster graph types, and is outperformed by the compositional methods for scale-free graph type in the low sample size regime.

In conclusion, SPRING exhibits considerably better graph recovery performance than existing methods, and is robust to misspecification of total sample abundance. This suggests that incorporating quantitative abundance information in the analysis leads to more reliable graphical model inference.

4. STATISTICAL MICROBIAL ASSOCIATIONS IN GUT MICROBIOME DATA

We provide two applications of SPRING to TAS-based microbial abundance data: a subset of the relative abundance data from the American Gut Project (AGP) (McDonald et al., 2018) and the QMP data from Vandeputte et al. (2017).

4.1. Taxon-Taxon Associations From the American Gut Project Data

We first use SPRING to infer taxon-taxon associations from the relative abundance AGP data. After the pruning and filtering steps described in section 3.3.1, we arrive at $p = 481$ OTUs from $n = 6482$ samples. Prior to applying SPRING, we transform the compositions $\mathbf{X} \in \mathbb{S}^{n\times p}$ using the mclr_ϵ transform introduced in Equation (2). The minimum value of the mclr_0-transformed data across all samples is $z_{\min} = -4.8142$. To make all non-zero values strictly positive, we add an arbitrary constant $c = 1$ to $|z_{\min}|$ and use the shift $\varepsilon = |z_{\min}| + c = 5.8142$ in the final mclr_ε transform. We also consider SPIEC-EASI, MB, and invSparCC (see **Table 2**) for comparison. All four methods use the same parameterization for the regularization path and StARS model selection: 50 subsamples with the same seed number, subsampling ratio ($10\sqrt{n}/n = 0.1242$) and 50 tuning parameter values with the same ratio of the smallest to largest λ value ($\lambda_{\min}/\lambda_{\max} = 0.01$). For each method, $\lambda_{\max}$ is set to the maximum value of the off-diagonal elements of the respective correlation matrix. All computations were performed in R using the R packages `pulsar`, `SpiecEasi`, `huge`, and `mixedCCA`, respectively.

We report summary statistics of the estimated association networks for two StARS stability thresholds: 0.05 (the standard setting in `SpiecEasi`) and 0.1 (the standard setting in Liu et al., 2010) in **Table 4**. For both stability thresholds, the MB method estimates the sparsest networks with highest percentage of positive edges (PEP) while invSparCC estimates the densest networks with the lowest percentage of positive edges. SPRING and SPIEC-EASI's association networks have similar edge densities while SPRING has a considerably higher percentage of positive partial correlation edges.

To get a bird's eye view of the topologies of the different association networks we visualize the four different networks at StARS threshold 0.05 in **Figure 9A**. The force-directed layout of all networks follows the optimal layout of the SPRING network. At the selected StARS threshold, all networks have one connected component. The overall network structure suggests a dense core with two peripheral network modules, similar to previous analysis (Müller et al., 2016). The networks of the compositionally-adjusted methods SPIEC-EASI and invSparCC

connect the core and one of the modules by a large number of positive (shown in green) and negative (shown in red) associations. SPRING considerably sparsifies these connections, leaving only few positive and negative edges between the modules, and MB does not infer any negative associations. We assess the similarity among the estimated networks by analyzing their edge set overlap in **Figure 9B**. All methods share common core of 601 edges. As expected, SPIEC-EASI and invSparCC share the largest unique two-set overlap with 637. SPRING's network takes an intermediate role between MB and the compositionally-adjusted methods. It shares 833 edges with SPIEC-EASI and invSparCC, and 112 edges exclusively with MB. Each method by itself also comprises a considerable set of exclusive edges, ranging from 418 for SPIEC-EASI to 767 for SPRING.

4.2. Genus-Genus Associations From Quantitative Gut Microbiome Profiling Data

We next analyze the quantitative gut microbiome data put forward in Vandeputte et al. (2017). We focus on estimating genus-genus associations both from the quantitative and the relative microbiome profiles, referred to as QMP and RMP, and analyze the consistency among the inferred networks. We follow the processing steps outlined in section 3.2.1 leading to $n = 106$ subjects and $p = 91$ genera. To infer statistical genus-genus associations we use SPRING for the QMP data (without transformation), and SPIEC-EASI for the corresponding RMP data (using the standard clr transformation) with the same computational protocol as detailed in the previous section.

We first show the agreement of signed edges between the two association networks at StARS stability level 0.1 in **Table 5**. Overall, out of the 4095 possible genus-genus associations, SPRING infers a set of 237 stable edges with a PEP of 98%. SPIEC-EASI infers 220 edges with a PEP of 66%. From the quantitative data, SPRING is able to detect considerably more positive associations, 140 of which are missed by SPIEC-EASI from the relative abundance data. SPRING detects only four negative associations three of which are missed by SPIEC-EASI despite having a considerable larger set of negative edges (74 overall). However, both methods do agree on a set of 93 edges, 92 positive and one negative edge. Importantly, we do not observe any sign flips among the different inferred edge sets. Missed positive or negative edges are simply absent in the other method.

We next focus on the induced genus-genus sub-network which only includes genera that have an assigned taxonomy and have at least one strong association $\geq |0.2|$ in either the SPRING-inferred or SPIEC-EASI-inferred association network. The weighted adjacency of this sub-network includes 32 genera and is shown in **Figure 10**. Among the 14 genera with highest total abundance across all samples (Bacteroides to Odoribacter), we observe 50% agreement between the two estimated networks (six edges are the same across all networks, three edges are different in SPIEC-EASI, four are different in SPRING). Both networks include a strong negative association between Phascolarctobacterium and Dialister and exactly four positive associations of Bacteroides with Parabacteroides, Holdemania, Bilophila, and Odoribacter (first row and column in **Figure 10**). We also observe the absence of a negative association between Bacteroides and Prevotella genera in the quantitative data which is often reported in the literature and also present in the SPIEC-EASI network (see also Vandeputte et al., 2017 for a discussion).

5. DISCUSSION

Advances in experimental microbiome profiling protocols have combined high-throughput environmental sequencing techniques with robust measurements of microbial cell counts

TABLE 2 | Summary of methods considered for comparison.

Method	Type of data	transformation	Correlation estimation
MB	Absolute abundance	None	Sample correlation
SPIEC-EASI	Relative abundance	clr	Sample correlation
invSparCC	Relative abundance	log	SparCC
SPRING	Absolute/relative abundance*	None/mclr*	SPR correlation

*For all methods, the final graphical model is estimated based on combining neighborhood selection approach with pulsar tuning parameter selection. *When absolute abundance data is not available, SPRING can be applied to relative abundance data following mclr transform described in section 2.3.*

TABLE 4 | AGP data: total number of partial correlation edges and percentage of positive partial correlation edges (PEP) (Faust et al., 2015) as estimated by MB, SPRING, SPIEC-EASI, and invSparCC for StARS stability thresholds $t_S = 0.05$ and 0.1.

	MB	SPRING	SPIEC-EASI	invSparCC
StARS threshold, t_S	Number of stable edges			
0.05	1621	2725	2702	3099
0.1	2970	4004	4008	4681
StARS threshold, t_S	Percentage of positive edges (PEP)			
0.05	1.0000	0.9835	0.8531	0.8341
0.1	0.9798	0.9515	0.7867	0.7584

TABLE 3 | Area under the Precision-Recall curves (AUPR) of **Figure 8**.

Dimension (n, p)	Graph type	SPIEC-EASI	SPRING	MB	invSparCC
(2000, 100)	Band	0.91 (0.92)	0.95 (0.94)	0.42 (0.34)	0.91 (0.91)
	Cluster	0.93 (0.93)	0.95 (0.92)	0.32 (0.27)	0.93 (0.93)
	Scale-free	0.93 (0.93)	0.96 (0.93)	0.26 (0.18)	0.94 (0.94)
(500, 200)	Band	0.89 (0.90)	0.93 (0.89)	0.20 (0.16)	0.87 (0.88)
	Cluster	0.83 (0.84)	0.90 (0.88)	0.25 (0.22)	0.81 (0.82)
	Scale-free	0.55 (0.54)	0.58 (0.50)	0.01 (0.01)	0.54 (0.54)

In each cell, AUPR of the True data and the Distorted data (given in parenthesis) are reported. AUPR value is based on edge selection probabilities using StARS with $t_S = 0.1$.

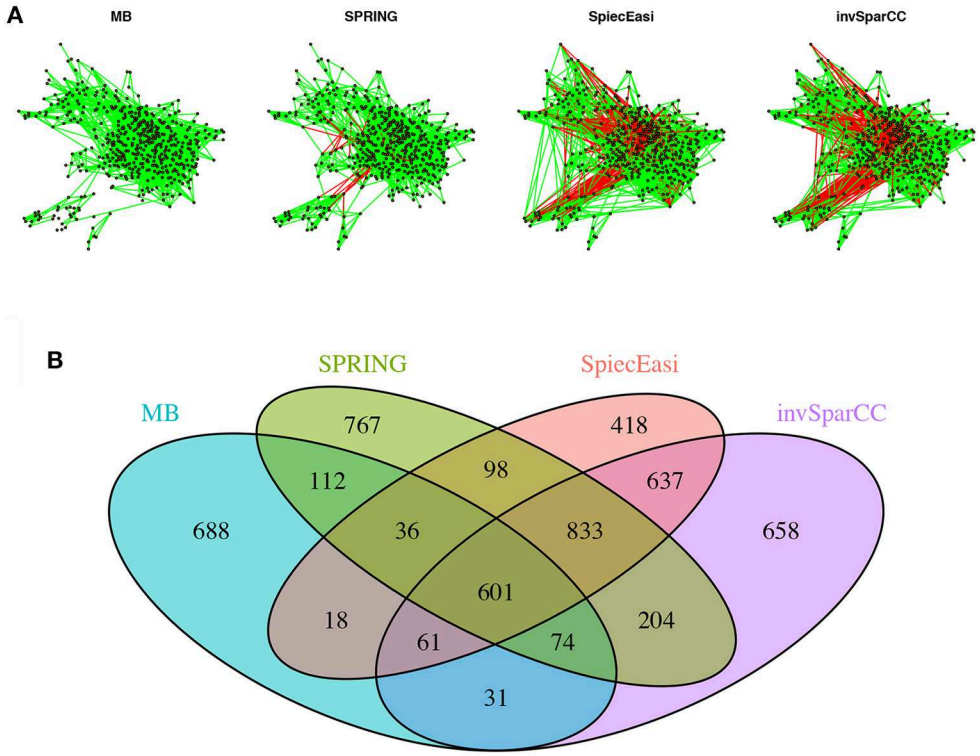


FIGURE 9 | Analysis of AGP data. **(A)** Force-directed layout (using `igraph` package in R) of the microbial association networks inferred by MB, SPRING, SPIEC-EASI, and invSparCC at StARS stability level 0.05. Green edges represent positive associations, red edges negative ones. **(B)** The four set Venn diagram showing the overlap of edge sets of the different association networks.

TABLE 5 | QMP data: summary of agreement of signed genus-genus partial correlations, inferred by SPRING and SPIEC-EASI at StARS stability threshold $t_S = 0.1$.

		SPRING		
Sign of estimated edges		**Positive**	**Zero**	**Negative**
SPIEC-EASI	Positive	92	54	0
	Zero	140	3731	4
	Negative	0	73	1

(Gifford et al., 2011; Satinsky et al., 2013; Stämmler et al., 2016; Props et al., 2017; Vandeputte et al., 2017; Tkacz et al., 2018), providing, for the first time, a more quantitative picture of the underlying microbial ecosystems in their natural habitat. To facilitate a high-level summary of the complex interplay between the constituents of the ecosystem, an important first exploratory analysis step is the estimation of statistical association networks between the identified operational taxonomic units or gene sets (Faust and Raes, 2012; Fuhrman et al., 2015; Sunagawa et al., 2015; Ruiz et al., 2017). In order to learn such association networks from sparse quantitative microbiome data, we have introduced the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING combines neighborhood selection (Meinshausen and Bühlmann, 2006) to infer the conditional dependency graph with stability-based model selection (Liu et al., 2010; Müller et al., 2016) to identify a sparse set of partial correlation estimates. The resulting network of partial correlations represents direct (i.e., conditionally independent) microbe-microbe associations and provides a statistical community-level description of the underlying microbial ecosystem. As ground truth microbial association networks are largely elusive in the literature, we have based our numerical simulation benchmarks on a novel synthetic quantitative microbiome data generation mechanism which might be of independent interest to researchers who want to test novel statistical techniques on such data.

Our benchmark test cases revealed a number of interesting observations. Firstly, we showed that, on synthetic quantitative microbiome data with prescribed ground-truth correlation structure, the SPR-type correlation estimates are considerably more accurate than SparCC and naive Pearson sample correlation on clr-transformed compositional data. Secondly, we showed that Pearson sample correlation estimation cannot be used to identify sparse partial correlations in quantitative microbiome data. Thirdly, SPRING outperformed sparse graphical modeling techniques that were designed with compositional data in mind, namely SPIEC-EASI (Kurtz et al., 2015) and the invSparCC estimator introduced here, which uses neighborhood selection with SparCC correlation estimation (Friedman and Alm, 2012). SPRING compared favorably to the other methods both in terms of achievability,

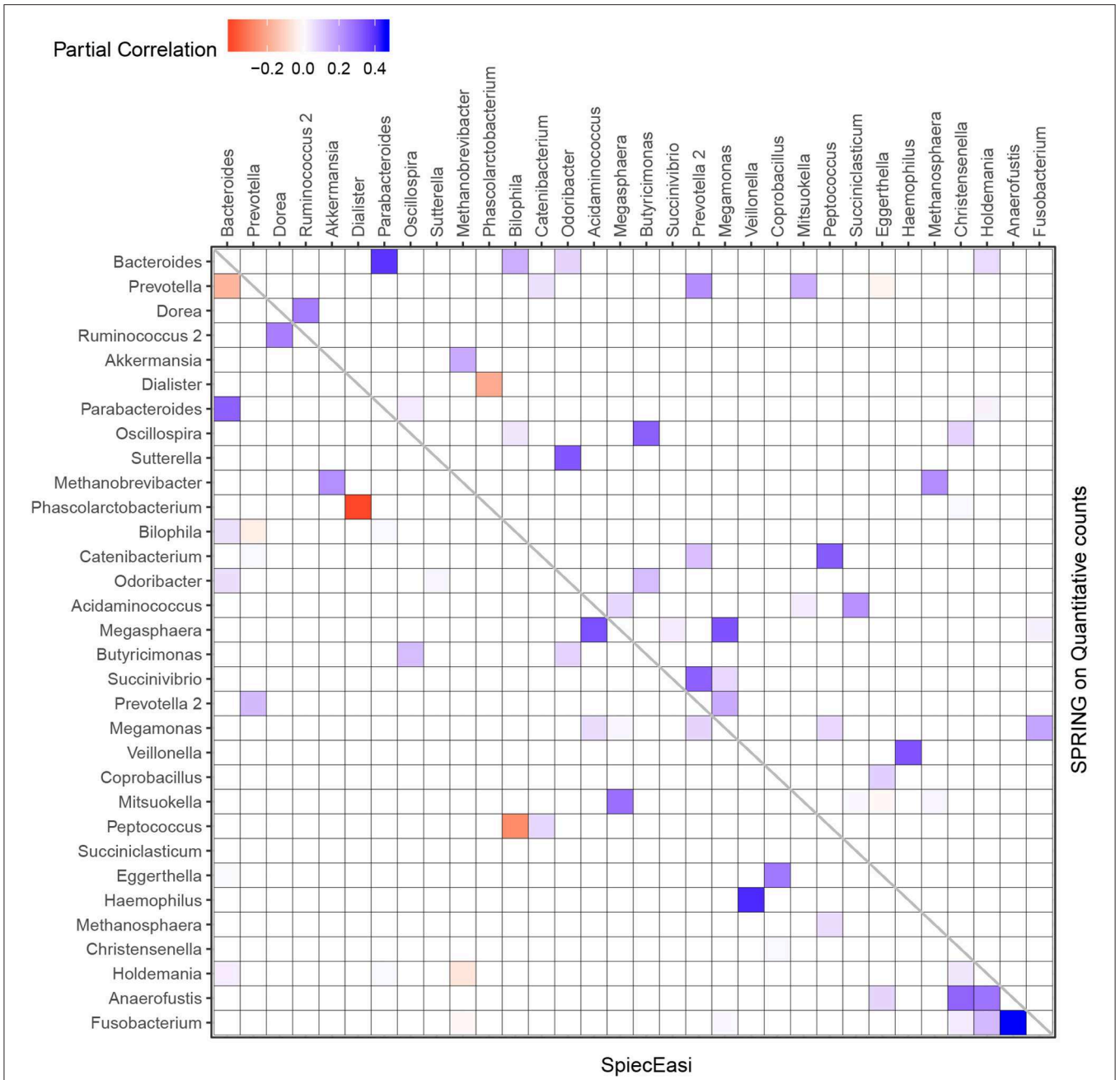


FIGURE 10 | Genus-genus association network using relative (lower triangular part) vs. quantitative count data (upper triangular part). Only genera with at least one strong association ≥ |0.2| in either SPIEC-EASI or SPRING are shown. The genera are ordered by the total quantitative abundance over healthy subjects ($n = 106$).

that is, in terms of minimum Hamming distance to the true underlying network achieved across the regularization path (see **Figure 7**), and in combination with stability-based model selection in terms of Precision-Recall (see **Figure 8**). We also quantified the robustness of SPRING to misspecification of the total by randomly distorting the counts of each sample up to a 6-fold change which represents a realistic distortion scenario in gut microbiome samples (see e.g., in Vandeputte et al., 2017, **Figure 2**) and is on the same order as typical fold changes of observed image-based total species counts in marine ecosystems (Ducklow, 2000). Even under these distortions SPRING's performance was on par or superior to SPIEC-EASI and invSparCC (which are scale-invariant by design). SPRING's robustness to total count misspecifications thus suggested to include an application of association inference from relative microbiome profiling data. In order to apply

SPRING to relative abundance data we introduced a modified centered log-ratio (clr) transform that can seamlessly handle excess zeros without pseudo-count addition. Contrary to recent efforts in data-driven pseudo-count inference (see de la Cruz and Kreft, 2018 and references therein) we computed the geometric mean of each sample from positive proportions only, normalized and log-transformed all non-zero proportions by using that geometric mean, and applied an identical shift operation to all non-zero variables in the dataset. This transformation is rank-preserving while leaving the original zero proportions unchanged, thus enabling the application of the SPRING methodology without further modification to relative abundance data.

We applied SPRING to two prominent gut microbial datasets, the relative abundance data collected in the American Gut Project (AGP) (McDonald et al., 2018) and the quantitative gut microbiome profiling (QMP) data from Vandeputte et al. (2017). As the processed data from Vandeputte et al. (2017) was not publicly available, a reprocessing of the amplicon sequencing reads was necessary.

From the AGP data, we inferred taxon-taxon association networks across $p = 481$ taxa from $n = 6482$ samples using neighborhood selection (MB), SPIEC-EASI, invSparCC, and SPRING. In line with previous findings (Faust et al., 2015), the percentage of positive edges in the networks is $> 75\%$, with MB and SPRING having even higher percentages than SPIEC-EASI and invSparCC. At both StARS stability levels 0.05 and 0.1 reported here, SPRING and MB tended to infer slightly sparser association networks than SPIEC-EASI and invSparCC. At StARS stability level 0.05, we analyzed the overlap of edge sets among the different methods (**Figure 9**). All methods share a common core of 601 edges. In addition, SPRING, SPIEC-EASI, and invSparCC shared the largest common edge set of size 833 among all three-set overlaps. As expected, the two compositionally-adjusted methods SPIEC-EASI and invSparCC shared the largest common two-set overlap of 637 edges. In the absence of verified taxon-taxon associations, our analysis suggests that a practitioner screening for coherent statistical associations among taxa can apply SPRING, SPIEC-EASI, and invSparCC independently and select the set of strongest edges out of the edge set these three methods inferred. This strategy is also supported by our synthetic benchmark results where the joint edge set of the three methods achieved a true positive rate of 0.95 for cluster graphs. For the analysis on the AGP data, this strategy would result in an edge set of size 1434, an average of about three associations per taxon. This core network can then be further studied in terms of modularity, network stability, and node centrality measures, as shown, e.g., in Ruiz et al. (2017); (Tipton et al., 2018).

For the QMP data, we used SPRING and SPIEC-EASI to estimate the genus-genus associations from the quantitative and the relative microbiome profiles, respectively. Our analysis revealed considerable differences to the published results in Vandeputte et al. (2017). The original study described dramatic differences between significant marginal genus-genus correlations from 66 healthy control samples in the QMP disease cohort when applying Spearman's ρ correlation to the relative and quantitative microbiome profiling data (see e.g., **Figure 3** in Vandeputte et al., 2017). Our results here showed more coherence of the statistical associations inferred from relative and absolute abundance data. Overall, 92 positive, 1 negative, as well as 3731 zero associations were in common among both association networks, while both networks differed in 280 associations (**Table 5**). Our analysis on the genus sub-network that comprised all genera with at least one strong association $\geq |0.2|$, shown in **Figure 10**, verified a strong negative association between Phascolarctobacterium and Dialister inferred from both data types, as well as the absence of a negative association between Bacteroides and Prevotella genera in the quantitative data, both in agreement with published results. However, we recovered, for both data types, exactly four positive associations for Bacteroides, namely with Parabacteroides, Holdemania, Bilophila, and Odoribacter (First row and column in **Figure 10**). The latter two associations were previously reported only to be present in the quantitative data. Overall, more than 30% of the edges in the sub-network agreed which is in marked contrast to the results reported in Vandeputte et al. (2017). The higher network consistency reported here can be attributed to several factors. Firstly, our amplicon data processing framework may result in slight differences in terms of OTU picking and avoids a rarefaction step which was included previously. Secondly, we considered partial rather than marginal correlations among the genera to avoid any influence of indirect associations. Thirdly, we analyzed both data types within the same coherent statistical learning framework: sparse learning of partial correlations via neighborhood selection followed by stability-based model selection with the identical stability threshold (here 0.1). Finally, we considered a larger sample size of $n = 106$ representing healthy subjects from two different cohorts available in the QMP data as opposed to the $n = 66$ samples used in the original study. We conclude that differences in association networks from relative and absolute abundance data are not only attributable to the data themselves but also highly method-dependent.

In summary, we believe that, as quantitative microbiome profiling will become increasingly available, the semi-parametric rank-based estimators for correlation and partial correlation estimation discussed here provide an important tool for reliable statistical analysis of quantitative microbiome data. While we have focused here on targeted amplicon-based sequencing datasets, our methodology is broadly applicable to other biological high-throughput data with large excess of zero counts, including quantitative metagenomics (Satinsky et al., 2013), single-cell RNA-Seq data (see Risso et al., 2018 for a recent statistical analysis framework), and mass spectrometry proteomics data (Drew et al., 2017). Moreover, the concept of SPR-type correlation employed in SPRING can naturally generalize to joint analysis of multi-omics dataset when, on the same sample, several zero-inflated data types are measured in tandem. The approach in Yoon et al. (2018) already exploits this idea for RNA-seq and micro-RNA data in the context of canonical correlation analysis. Extending SPRING in a similar way to joint graphical modeling of mixed data types is a promising next step toward a consistent

and coherent statistical analysis framework for sparse high-throughput biological datasets.

AUTHOR CONTRIBUTIONS

GY, IG, and CM developed the methodology. GY lead the numerical analysis. IG assisted GY in simulation studies. CM assisted GY in data analyses. GY, IG, and CM prepared the manuscript.

FUNDING

GY was supported by the National Cancer Institute grant T32-CA090301. IG was supported by the National Science Foundation grant DMS-1712943. CM work was supported by the Simons Foundation.

ACKNOWLEDGMENTS

We are grateful to Dr. Zachary D. Kurtz, Lodo Therapeutics, for providing the processed American Gut data and the QMP data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00516/full#supplementary-material

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65.

Aitchison, J. (2003). "A concise guide to compositional data analysis," in *2nd Compositional Data Analysis Workshop* (Girona).

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., et al. (2018). Structure and function of the global topsoil microbiome. *Nature* 560, 233–237. doi: 10.1038/s41586-018-0386-6

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

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581. doi: 10.1038/nmeth.3869

Cao, Y., Lin, W., and Li, H. (2018). Large covariance estimation for compositional data via composition-adjusted thresholding. *J. Am. Stat. Assoc.* 1–14. doi: 10.1080/01621459.2018.1442340

Cao, Y., Zhang, A., and Li, H. (2017). Microbial composition estimation from sparse count data. *ArXiv e-prints*. Available online at: http://arxiv.org/abs/1706.02380 (accessed May 27, 2019).

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. *Nature* 7, 335–336. doi: 10.1038/nmeth.f.303

de la Cruz, R., and Kreft, J.-U. (2018). Geometric mean extension for data sets with zeros. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1806.06403 (accessed May 27, 2019).

Drew, K., Müller, C. L., Bonneau, R., and Marcotte, E. M. (2017). Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. *PLoS Comput. Biol.*, 13:e1005625. doi: 10.1371/journal.pcbi.1005625

Ducklow, H. W. (2000). "Bacterial production and biomass in the oceans," in *Microbial Ecology of the Oceans*, ed D. L. Kirchman (New York, NY: Wiley-Liss), 85–120.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. Available online at: https://www.biorxiv.org/content/early/2016/10/15/081257 (accessed May 27, 2019).

Egozcue, J. J., Pawlowsky-Glahn, V., and Gloor, G. B. (2018). Linear association in compositional data analysis. *Aust. J. Stat.* 47:3. doi: 10.17713/ajs.v47i1.689

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. B* 79, 405–421. doi: 10.1111/rssb.12168

Faust, K., Lima-Mendez, G., Lerat, J.-S., Sathirapongsasuti, J. F., Knight, R., Huttenhower, C., et al. (2015). Cross-biome comparison of microbial association networks. *Front. Microbiol.* 6:1200. doi: 10.3389/fmicb.2015.01200

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

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Fuhrman, J., Cram, J., and Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* 13, 133–146. doi: 10.1038/nrmicro3417

Gifford, S. M., Sharma, S., Rinta-Kanto, J. M., and Moran, M. A. (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J.* 5, 461–472. doi: 10.1038/ismej.2010.141

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

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

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kurtz, Z., Mueller, C., Miraldi, E., and Bonneau, R. (2017). *SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference*. R package version 1.0.2.

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

Lagkouvardos, I., Fischer, S., Kumar, N., and Clavel, T. (2017). Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* 5:e2836. doi: 10.7717/peerj.2836

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Li, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi: 10.1093/biomet/asu031

Liu, H., Han, F., Yuan, M., Lafferty, J. D., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.* 40, 2293–2326. doi: 10.1214/12-AOS1037

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* 10, 2295–2328. doi: 10.1145/1577069.1755863

Liu, H., Roeder, K., and Wasserman, L. (2010). "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)* (Vancouver, BC), 1432–1440.

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, 1–7. doi: 10.3402/mehd.v26.27663

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. doi: 10.1128/mSystems.00031-18

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34, 1436–1462. doi: 10.1214/009053606000000281

Müller, C. L., Bonneau, R., and Kurtz, Z. (2016). Generalized stability approach for regularized graphical models. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1605.07072 (accessed May 27, 2019).

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, 584–587. doi: 10.1038/ismej.2016.117

Quinn, T. P., Richardson, M. F., Lovell, D., and Crowley, T. M. (2017). Propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* 7, 1–9. doi: 10.1038/s41598-017-16520-0

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284. doi: 10.1038/s41467-017-02554-5

Ruiz, V. E., Battaglia, T., Kurtz, Z. D., Bijnens, L., Ou, A., Engstrand, I., et al. (2017). A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity. *Nat. Commun.* 8:518. doi: 10.1038/s41467-017-00531-6

Satinsky, B. M., Gifford, S. M., Crump, B. C., and Moran, M. A. (2013). Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol.* 531, 237–250. doi: 10.1016/B978-0-12-407863-5.00012-5

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical assessment of metagenome interpretation–a benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071. doi: 10.1038/nmeth.4458

Sedlar, K., Kupkova, K., and Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* 15, 48–55. doi: 10.1016/j.csbj.2016.11.005

Soetaert, K. (2009). *rootSolve: Nonlinear Root Finding, Equilibrium and Steady-State Analysis of Ordinary Differential Equations*. R package version 1.6.

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., et al. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 4:28. doi: 10.1186/s40168-016-0175-0

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., et al. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* 6:12. doi: 10.1186/s40168-017-0393-0

Tkacz, A., Hortala, M., and Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* 6, 1–13. doi: 10.1186/s40168-018-0491-7

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511. doi: 10.1038/nature24460

Woese, C., and Fox, G. (1977). Phylogenetic structure of the prokaryotic domain. *PNAS* 74, 5088–5090.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *J. Stat. Softw.* 32, 1–34. doi: 10.18637/jss.v032.i10

Yoon, G., Carroll, R. J., and Gaynanova, I. (2018). Sparse semiparametric canonical correlation analysis for data of mixed types. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1807.05274 (accessed May 27, 2019).

Yoon, G., and Gaynanova, I. (2018). *mixedCCA: Sparse CCA for High-Dimensional Mixed Data*. R package version 1.0.1.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13, 1059–1062. Available online at: http://www.jmlr.org/papers/v13/zhao12a.html

Conflict of Interest Statement: 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

REVIEW
published: 25 June 2019
doi: 10.3389/fgene.2019.00579

A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction

Yi-Hui Zhou[1]* *and Paul Gallins*[2]

[1] Department of Biological Sciences, North Carolina State University, Raleigh, NC, United States, [2] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, United States

With the growing importance of microbiome research, there is increasing evidence that host variation in microbial communities is associated with overall host health. Advancement in genetic sequencing methods for microbiomes has coincided with improvements in machine learning, with important implications for disease risk prediction in humans. One aspect specific to microbiome prediction is the use of taxonomy-informed feature selection. In this review for non-experts, we explore the most commonly used machine learning methods, and evaluate their prediction accuracy as applied to microbiome host trait prediction. Methods are described at an introductory level, and R/Python code for the analyses is provided.

Keywords: disease, phenotype, modeling, machine learning, prediction

OPEN ACCESS

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

Reviewed by:
Himel Mallick,
Merck, United States
Jun Chen,
Mayo Clinic, United States

***Correspondence:**
Yi-Hui Zhou
yihui_zhou@ncsu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 January 2019*
Accepted: *04 June 2019*
Published: *25 June 2019*

Citation:
Zhou Y-H and Gallins P (2019) A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front. Genet. 10:579. doi: 10.3389/fgene.2019.00579

1. INTRODUCTION

The microbiome is the collection of all microbes living in or on a host, including bacteria, viruses, and fungi (Robinson and Pfeiffer, 2014). The risk or severity of numerous diseases and disorders in a host are associated with the microbiome (Kinross et al., 2011), and accurate trait prediction based on microbiome characteristics is an important problem (Rothschild et al., 2018). The application of modern machine learning algorithms is proving to be valuable in this effort (Gilbert et al., 2018). This review/tutorial focuses on the bacterial component of the microbiome, although in principle many of the elements apply more generally.

With modern high-throughput sequencing, entire microbial communities can be profiled, revealing an extensive diversity of genes and organisms (Turnbaugh et al., 2007). A common strategy is to sequence only a highly specific region, such as 16S ribosomal RNA (rRNA), although the methods described below can also be applied to metagenomic shotgun methods (Mande et al., 2012). Due to the graded nature of sequence similarity, the data are often organized into operational taxonomic units (OTUs) (Schmitt et al., 2012), i.e., clusters of similar sequences, intended to represent the abundance of a particular bacterial taxon while avoiding excessive sparsity that would result if only identical sequences were grouped. Typical choices of similarity limits (e.g., grouping sequences with no more than 3% dissimilarity) produce taxa that are specific to bacterial species, or represent a further subdivision within species. Informatic methods for taxonomic classification use databases (McDonald et al., 2012), such as SILVA (Quast et al., 2012), and are beyond our scope, but we assume that such classification is available. The result after OTU grouping is a matrix (OTU table) of OTU features by the number of samples, where the number of features can vary dramatically across datasets due to stringency of grouping. Although methods that avoid OTU grouping have been described (Callahan et al., 2016), OTU tables remain common and are a practical starting point for most machine learning prediction methods. For additional discussion

of levels of taxonomy, with intriguing thoughts about the interplay and use of molecular function descriptors vs. taxonomic descriptors, the reader is referred to Knights et al. (2011b) and Xu et al. (2014). However, many of the principles discussed here apply regardless of the feature type.

Several features of OTU tables present challenges. First, OTU tables are sparse, with a large proportion of zero counts (Hu et al., 2018). Investigators have often removed OTUs that were present in too few samples to be useful, or collapsed OTUs into the genus level, which is a simple form of "feature engineering" that we will explore further below. Second, the role of taxonomy in prediction is often unclear – similar sequences are often correlated across samples, which is a property that can be readily assessed directly without taxonomic knowledge. Third, as with many omics technologies, library sizes (essentially column sums of the OTU table) vary considerably, and normalization methods must be used to account for this variation (Weiss et al., 2017).

A number of excellent reviews have been published, covering experimental design and targeted amplicon vs. metagenomics profiling (Mallick et al., 2017), and a comprehensive overview of different experimental and interrogation methods and analyses (Knight et al., 2018). Other reviews have covered the remarkable advances in understanding that have resulted recently in understanding connections of, e.g., human gut microbiome populations to human health (Cani, 2018).

Recently, studies have begun to explore the power of machine learning to use microbiome patterns to predict host characteristics (Knights et al., 2011a; Moitinho-Silva et al., 2017). Existing studies often report disease-associated dysbiosis, a microbial imbalance inside the host, but such associations can have a wide range of interpretations. Individual studies have also suffered from small sample sizes, inconsistent findings, and a lack of standard processing and analysis methods (Duvallet et al., 2017). Prediction models have sometimes been difficult to generalize across studies (Pasolli et al., 2016). One approach to resolve these issues is by performing a meta-analysis, combining microbiome studies across common traits. Duvallet et al. (2017) have performed a cross-disease meta-analysis of published case-control gut microbiome studies spanning 10 diseases. They found consistent patterns characterizing disease-associated microbiome changes and concluded that many associations found in case-control studies are likely not disease-specific but rather part of a non-specific, shared response to health and disease. Pasolli et al. (2016) also performed a meta-analysis in a collection of 2,424 publicly available samples from eight large-scale studies. The authors remarked that addition of healthy (control) samples from other studies to training sets improved disease prediction capabilities. Nonetheless, any meta- or pooled analysis should rely on a solid foundation of effective per-study prediction. The use of multiple studies enabled Pasolli et al. (2016) to explore the use of external validation of models across truly separate datasets. Such external validation can in principle result in more robust and generalizable models for prediction than models that are validated internally only.

Sophisticated machine learning methods in microbiome analysis have been proposed considerably in recent years, including using deep neural networks (Ananthakrishnan et al., 2017), and leveraging methods for genomes and metagenomes (Rahman et al., 2018). However, the content-knowledge required to implement these methods is high, presenting a barrier to data scientists looking to get started in microbiome analysis and prediction. Moreover, there are few resources for biologists with intermediate statistical and computing background to "jump in" to analysis of the important trait prediction problem. The target audience of this paper is those seeking a brief review and tutorial for trait prediction, and who will benefit from accessible code. After digesting these basic building blocks of analysis, the reader may move to more advanced, such as dynamic systems modeling (Brooks et al., 2017).

The remainder of this paper is written in several sections. Section 2 reviews the steps of data preparation before machine learning implementation. Section 3 provides a quick overview of the most commonly-used machine learning (ML) methods, as well as the most commonly used performance criteria. Experienced modelers can skip this section. Section 4 summarizes the scope of the relevant literature and describes several real datasets and the trait of interest. Section 5 provides results, and the underlying code forms a tutorial of machine learning methods applied in this context.

2. DATA PREPARATION

Many machine learning methods have difficulty with missing features, and so we assume the OTU table is complete. A minor fraction of missing data can often be effectively handled using simple imputation procedures, such as kNN-impute (Crookston and Finley, 2008), or even simpler methods, such as feature-median imputation. The methods described in this section, including imputation and normalization, must be performed without using the host trait information, because otherwise they might be biased by this information. Feature selection methods that use host trait information belong in the next section, as they must be included inside a cross-validation procedure.

2.1. Notation and Sampling Considerations

Let X be an $m \times n$ matrix of microbiome count data, where m is the number of OTU features and n is the number of samples. Let y be a vector of length n with the microbiome host trait. Commonly a trait will be a binary outcome (e.g., case/control status, coded 1/0), or a continuous trait, such as body mass index (BMI). Here our use of microbiome features as predictive of a trait does not imply or assume causality. We note that case/control study designs often involve oversampling of one type (often cases) relative to the general population. A prediction rule might explicitly use this information, for example by a simple application of Bayes' rule (Tibshirani et al., 2003), with prior probabilities reflecting those in the general population. Such sampling considerations are beyond our scope, and we refer the reader to Chawla (2009). Here we consider our sample dataset to be representative of the population of its intended downstream use.

2.2. Transformation and Normalization

Normalization is an essential process to ensure comparability of data across samples (Weiss et al., 2017), largely to account for the large variability in library sizes (total number of sequencing reads across different samples). The basic issues are similar to those encountered in expression sequence normalization (de Kok et al., 2005), but less is currently known about sources of potential bias to inform microbiome normalization. Normalization methods assessed by Weiss et al. (2017) included cumulative sum scaling, variance stabilization, and trimmed-mean by *M*-values. Randolph et al. (2018) utilized the centered log-ratio (CLR) transform of the relative abundance vectors, based on a method developed by Aitchison (1982), replacing zeros with a small positive value. As part of their motivation, Randolph et al. (2018) pointed out that standard cumulative sum scaling places the normalized data vectors in a simplex, with potential consequences for kernel-based discovery methods (Randolph et al., 2018).

2.3. Taxonomy as Annotation

Taxonomy is the science of defining and naming groups of biological organisms on the basis of shared characteristics. In our context, taxonomy refers to the evolutionary relationship among the microbes represented by each OTU, from general to specific: kingdom, phylum, class, order, family, genus and species, and OTU (Oudah and Henschel, 2018). For example, Kostic et al. (2012) summarized their findings in the study of microbiota in colorectal cancer using genera and phyla-level summaries, illustrating the importance of taxonomy in interpretation. Here we are highlighting the use of taxonomy in *post-hoc* interpretation of findings, providing important biological context. However, if the taxonomy is used in a supervised manner to improve prediction, it then becomes part of the formal machine learning procedure, as described in the next section.

3. REVIEW OF MACHINE LEARNING METHODS FOR PREDICTION

Machine learning deals with the creation and evaluation of algorithms to recognize, classify, and predict patterns from data (Tarca et al., 2007). Unsupervised methods identify patterns apparent in the data, but without the use of pre-defined labels (traits, in our context). These methods include (i) hierarchical clustering, which builds a hierarchy of clusters using a dendrogram, combining or splitting clusters based on a measure of dissimilarity between vectors of X; and (ii) k-means clustering, which involves partitioning the n vectors of X into k clusters in which each observation is classified to a cluster mean according to a distance metric. Unsupervised methods are important exploratory tools to examine the data and to determine important data structures and correlation patterns.

For the host trait prediction problem, we focus on supervised methods, in which labels (traits) of a dataset are known, and we wish to train a model to recognize feature characteristics associated with the trait. A primary difficulty in the problem is that the number of features (m rows) in the OTU table may greatly exceed the sample size n, so that over-fitting of complex models to the data is a concern.

3.1. Training and Cross-Validation

Training a model in supervised learning amounts to finding a parameter vector β that represents a rule for predicting a trait y from an m-vector x. This rule may take the form of a regression equation or other prediction rule. Prediction rules that use only a few features (n or fewer) are referred to as "sparse." A good prediction rule has high accuracy, as measured by quantities, such as the area under the receiver-operator characteristic curve, or the prediction correlation R, both described below. Many prediction methods proceed by minimizing an objective function $obj(\beta) = L(\beta) + \Omega(\beta)$, which contains two parts: the raw training loss L and a regularization term Ω. The training loss measures how predictive the model is with respect to the data used to train the model, and the regularization term penalizes for the complexity of the model, which helps to avoid overfitting.

An essential component of machine learning is the use of cross-validation to evaluate prediction performance, and often to select tuning parameters that govern the complexity of the model. One round of k-fold cross-validation involves partitioning the n samples into k subsets of roughly equal size, using each subset in turn as as the validation data for testing the algorithm, with the remaining samples as the training set. After a single round of cross-validation, each sample i has an associated predicted trait value $\hat{y}_i$, where the prediction rule was developed without any knowledge of the data from sample i (or at least without knowledge of y_i). The performance measure is computed by comparing the length-n $\hat{y}$ vector to the true y. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds to give an estimate of the predictive performance. Although the term "cross-validation" formally refers to the use of each sample i as both part of the training set and as testing set (i.e., crossing) during a single round, the term is often used more generically. For example, researchers sometimes use a simple holdout method in which a fraction $1/k$ of the data are randomly selected as a test set, the remainder as training, and repeat the process randomly with enough rounds to provide a stable estimate of accuracy.

3.2. Taxonomy and Structural Feature Extraction

Our Results section shows the results of prediction methods using all OTUs, as well as reduced-OTU selected or aggregated features. Several methods have been proposed to reduce the number of OTU features using correlation and taxonomy information, including Fizzy (Ditzler et al., 2015a), MetAML (Pasolli et al., 2016), and HFE (Oudah and Henschel, 2018). Aspects of the approaches are supervised and thus must be handled inside a cross-validation procedure.

For simplicity, here we focus on the hierarchical feature engineering (HFE) algorithm created by Oudah and Henschel (2018), which uses correlation and taxonomy information in X to exploit the underlying hierarchical structure of the feature space.

The HFE algorithm consists of four steps: (1) feature engineering: consider the relative abundances of higher taxonomic units as potential features by summing up the relative abundances of their respective children in a bottom-up tree traversal; (2) correlation-based filtering: calculate the correlation of values for each parent-child pair in the taxonomy hierarchy, and if the result is greater than a predefined threshold, then the child node is discarded; (3) information gain (IG) based filtering, reflecting association of features to the trait: construct all paths from the leaves (OTUs) to the root and for each path, calculate the IG of each node with respect to the trait values, and then calculate and use the average IG as a threshold to discard any node with a lower IG score; (4) IG-based leaf filtering: for OTUs with incomplete taxonomic information, discard any leaf with an IG score less than the global average IG score of the remaining nodes from the third phase. Steps (3) and (4) must be cross-validated, as they use the trait values. The python code for implementation is on our site (https://sites.google.com/ncsu.edu/zhouslab/home/software?).

The result is a set of informative features, perhaps including original OTUs along with higher-level aggregations of taxonomic features, that can be utilized for downstream machine learning (Oudah and Henschel, 2018). Standard feature selection algorithms, Fizzy and MetAML, which do not capitalize on the hierarchical structure of features, were also tested by Oudah and Henschel (2018) using several machine learning methods on real datasets. Since HFE was reported to outperform other methods (Oudah and Henschel, 2018) and resulted in higher prediction performance overall, we apply it in the real data analysis section to extract OTU features before applying machine learning methods of trait prediction. Note that feature selection can in principle be performed inside a grand cross-validation and prediction loop, or performed prior to prediction, as we have done for convenience here.

3.3. Supervised Learning Methods Commonly Used in Trait Prediction

Here we list the learning methods most commonly used in microbiome host trait prediction. The list is not exhaustive, but reflects our review of the methods in common use. In particular, neural networks have received considerable recent attention, but it is difficult to find quantitative evidence for the additional predictive ability in comparison to other methods. For several of the methods, it is common to center and row-scale X prior to application of the method, so each feature is given similar "weight" in the analysis.

3.3.1. Regression

The use of linear models enables simple fitting of continuous traits y as a function of feature vectors. However, if $m \geq n$ then structural overfitting occurs, and even if $m < n$ accuracy is often improved by using penalized (regularized) models. For the model $y = X\beta + \epsilon$, the training loss is $\sum_i (y_i - \hat{y}_i)^2$ the most commonly-used regularization methods are ridge regression (Hoerl and Kennard, 1970) and Lasso (Tibshirani, 1996) regression, which respectively use penalties $\lambda \sum_i \beta_i^2$ and $\lambda \sum_i |\beta_i|$ (not including the intercept) to the training loss. For binary class prediction, the approach is essentially the same, applying a generalized linear (logit) model, with the negative log-likelihood as the training loss. Here λ is a tuning parameter that can be optimized as part of cross-validation. Both methods provide "shrunken" coefficients, i.e., closer to zero than an ordinary least-squares approach. The results for Lasso are also sparse, with no more than n non-zero coefficients after optimization, and thus Lasso is also a feature-selection method. Another variant is the elastic net (Zou and Hastie, 2005), an intermediate version that linearly combines both penalties.

3.3.2. Linear Discriminant Analysis (LDA)

For binary traits, this approach finds a linear combination of OTUs in the training data that models the multivariate mean differences between classes (Lachenbruch and Goldstein, 1979). Classical LDA assumes that feature data arise from two different multivariate normal densities according to $y = 0$ and $y = 1$, i.e., $MVN(\mu_0, \Sigma)$ and $MVN(\mu_1, \Sigma)$ (**Figure 1A**). The prediction value is the estimate of the posterior mean $E(Y|x) = Pr(Y = 1|X)$, used because it minimizes mean-squared error.

3.3.3. Support Vector Machines (SVM)

This is another approach in the linear classifier category (**Figure 1A**), but in contrast to LDA may be considered non-parametric. In SVM, the goal is to find the hyperplane in a high-dimensional space that represents the largest margin between any two instances (support vectors) of two classes of training-data points, or that maximizes a related function if they cannot be separated. Non-linear versions of SVM are devised using a so-called kernel similarity function (Cortes and Vapnik, 1995).

3.3.4. Similarity Matrices and Related Kernel Methods

Some applications of microbiome association testing have compared similarity matrices across features to similarity of traits (Zhao and Shojaie, 2016). A closely-related approach is to first compute principal component (PC) scores, which may be obtained from OTU sample-sample correlation matrices (Zhou et al., 2018), and to use these PC scores as trait predictors. Kernel-penalized regression, an extension of PCA, was utilized by Randolph et al. (2018). in their microbiome data analysis. They applied a significance test for their graph-constrained estimation method, called Grace (Zhao and Shojaie, 2016), to test for association between microbiome species and their trait. However, trait prediction is not available in their software.

3.3.5. *k*-Nearest Neighbors (*k*-NN)

Training samples are vectors in a multi-dimensional space, each with a class label or continuous trait value. For discrete traits, a test sample is assigned the label which is most frequent among the k training samples nearest to that point (**Figure 1B**). Euclidean distance or correlation coefficients are the most commonly used distance metrics. For continuous traits, a weighted average of the k nearest neighbors is used, sometimes weighted (e.g., by the inverse of their distance from the new point).

3.3.6. Random Forests

Random forests (Breiman, 2001) are an increasingly used method, extensively applied in many different fields, including computational biology and genomics (Statnikov et al., 2013)

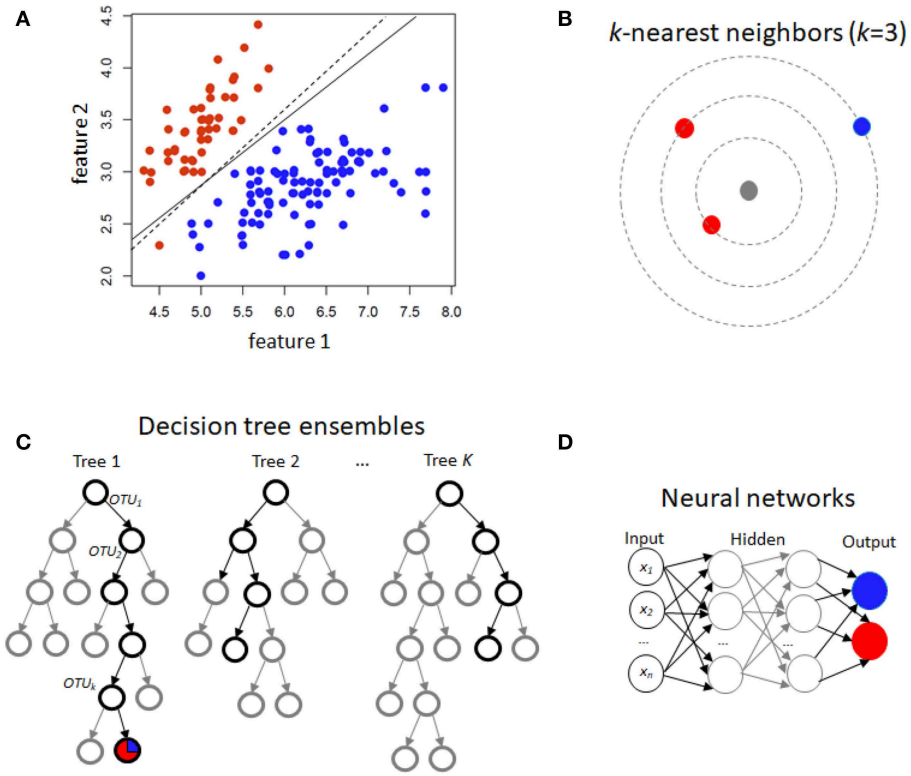


FIGURE 1 | Schematic illustration of several machine learning prediction methods using case/control (red/blue) status. For two features, **(A)** illustrates linear discrimination methods. The solid line shows the linear discriminant line corresponding to equally probable outcomes, while the dashed line shows the midpoint of the maximum-margin support vector machine. **(B)** For *k*-nearest neighbors, the gray point is predicted using an average of the neighbors (red, in this instance). **(C)** Decision tree ensembles include random forests, which average over bootstrapped trees, and boosted trees, where successive residuals are used for fitting. Trees may not extend to the level of individual observations, and modal or mean values in the terminal nodes are used for prediction. **(D)** A neural network with few hidden layers.

The building block of a "forest" is a decision tree, which uses features and associated threshold values to successively split the samples into groups that have similar y values. This process is repeated until the total number of specified nodes is reached. An ensemble of decision trees (or regression trees for continuous y) is built by performing bootstrapping on the dataset and averaging or taking the modal prediction from trees (a process known as "bagging")(**Figure 1C**), with subsampling of features used to reduce generalization error (Ho, 1995). An ancillary outcome of the bootstrapping procedure is that the data not sampled in each bootstrap (called "out of bag") can be used to estimate generalization error, as an alternative to cross-validation.

3.3.7. Gradient Boosting

Gradient boosting for decision trees refers to a process of ensemble modeling by averaging predictions over decision trees (learners) of fixed size (Friedman, 2001). As with other forms of boosting, the process successively computes weights for the individual learners in order to improve performance for the poorly-predicted samples. Following observations that boosting can be interpreted as a form of gradient descent on a loss function (such as $\sum_i (y_i - \hat{y}_i)^2$), gradient tree boosting successively fits decision trees on quantities known as "pseudo-residuals" (Friedman, 2002) for the loss function (**Figure 1C**).

3.3.8. Neural Networks

Neural networks refer to an interconnected feed-forward network of nodes ("neurons") with weights attached to each edge in the network, which allows the network to form a mapping between the inputs X and the outcomes y (Ditzler et al., 2015a). Each neuron j receiving an input $p_j(t)$ from predecessor neurons consists of the following components: an activation $a_j(t)$, a threshold θ_j, an activation function f that computes the new activation at a given time $t + 1$, and an output function f_{out} computing the output from the activation. These networks contain either one or many hidden layers, depending on the network type (**Figure 1D**). For microbiome data, the input layer is the set of OTUs, with separate neurons for each OTU. Hidden layers use backpropagation to optimize the weights of the input variables in order to improve the predictive power of the model. The total number of hidden layers and number of neurons within each hidden layer are specified by the user. All neurons from the input layer are connected to all neurons in the first hidden layer, with weights representing each connection. This process continues until the last hidden layer is connected

TABLE 1 | Review of published prediction accuracy comparisons.

Paper	Dataset	Trait	Samples	Cases	Controls	Taxa	Level	Method	Metric	Value
Pasolli et al., 2016	Qin et al., 2014	Liver cirrhosis	232	118	114	542	Species	Random forest	AUC	0.95
								SVM	AUC	0.92
								Elastic net	AUC	0.91
								Lasso	AUC	0.88
	Zeller et al., 2014	Colorectal cancer	121	48	73	503	Species	Random forest	AUC	0.87
								SVM	AUC	0.81
								Elastic net	AUC	0.79
								Lasso	AUC	0.73
	Qin et al., 2010	IBD	110	25	85	443	Species	Random forest	AUC	0.89
								SVM	AUC	0.86
								Elastic net	AUC	0.83
								Lasso	AUC	0.81
	Le Chatelier et al., 2013	Obesity	253	164	89	465	Species	Random forest	AUC	0.66
								SVM	AUC	0.65
								Elastic net	AUC	0.64
								Lasso	AUC	0.60
	Qin et al., 2012	Type II diabetes	344	170	174	572	Species	Random forest	AUC	0.74
								SVM	AUC	0.66
								Elastic net	AUC	0.70
								Lasso	AUC	0.71
	Karlsson et al., 2013	Type II diabetes	96	53	43	381	Species	Random forest	AUC	0.76
								SVM	AUC	0.66
								Elastic net	AUC	0.60
								Lasso	AUC	0.54
Johnson et al., 2016		Post-mortem interval (PMI)	67	NA	NA	52	Phylum	Ridge	Error rate	0.46
						52	Phylum	Elastic net	Error rate	0.48
						3,130	Species	Lasso	Error rate	0.49
						52	Phylum	SVM	Error rate	0.50
						3,130	Species	Ridge	Error rate	0.51
						3,130	Species	Elastic net	Error rate	0.52
						52	Phylum	Lasso	Error rate	0.52
Ditzler et al., 2015b	Rousk, 2010	Soil pH (low/medium/high)	22	NA	NA	500	Various	Recursive neural network (RNN) (50)	Error rate	0.15
								Deep belief network (DBN) (500)	Error rate	0.08
								Deep belief network (DBN) (750)	Error rate	0.08
								Random forest	Error rate	0.15
								Multi-layer perceptron Neural network (MLPNN) (500)	Error rate	0.00
	Caporaso et al., 2011	Host gender	1,967	NA	NA	500	various	Recursive neural network (RNN) (250)	Error rate	0.15
								Recursive neural network (RNN) (500)	Error rate	0.19
								Deep belief network (DBN) (250)	Error rate	0.24
								Deep belief network (DBN) (500)	Error rate	0.24

(Continued)

TABLE 1 | Continued

Paper	Dataset	Trait	Samples	Cases	Controls	Taxa	Level	Method	Metric	Value
								Random forest	Error rate	0.03
								Multi-layer perceptron neural network (MLPNN) (500)	Error rate	0.08
	Caporaso et al., 2011	Three body sites	1,967	NA	NA	500	Various	Recursive neural network (RNN) (250)	Error rate	0.17
								Recursive neural network (RNN) (500)	Error rate	0.16
								Deep belief network (DBN) (250)	Error rate	0.03
								Deep belief network (DBN) (500)	Error rate	0.03
								Random forest	Error rate	0.01
								Multi-layer perceptron neural network (MLPNN) (500)	Error rate	0.01
Reiman et al., 2017	Caporaso et al., 2011	Three body sites	1,967	NA	NA	1,706	Various	Recursive neural network (RNN) (250)	Accuracy	0.83
								Recursive neural network (RNN) (500)	Accuracy	0.84
								Deep belief network (DBN) (250)	Accuracy	0.97
								Deep belief network (DBN) (500)	Accuracy	0.97
								Multi-layer perceptron Neural network (MLPNN) (500)	Accuracy	0.99
								Random forest	Accuracy	0.99
								Convolutional neural Network (CNN-1D)	Accuracy	0.95
								Convolutional neural Network (CNN-2D)	Accuracy	0.99
Moitinho-Silva et al., 2017		Microbial abundance from sponges (high/low)	1,232	NA	NA	30	Phylum	random forest	Accuracy	0.97
								Adaptive boosting (AdaBoost)	Accuracy	0.95
						76	Class	Random forest	Accuracy	0.95
								Adaptive boosting (AdaBoost)	Accuracy	0.91
						2,322	Various	Random forest	Accuracy	0.50
								Adaptive boosting (AdaBoost)	Accuracy	0.91
Ai et al., 2017		Colorectal cancer (CRC)	141	42	99	1,171	Species	Bayes net	AUC	0.93
								Random forest	AUC	0.94
								Logistic	AUC	0.98
			141	53	88	783	Species	Bayes net	AUC	0.86
								Random forest	AUC	0.86
								Logistic	AUC	0.71
Wu et al., 2018		Three diseases	806	423	383	300	Genus	Logistic	F1	0.91
								k-nearest neighbor	F1	0.86
								Random forest	F1	0.83
								SVM	F1	0.91

(Continued)

TABLE 1 | Continued

Paper	Dataset	Trait	Samples	Cases	Controls	Taxa	Level	Method	Metric	Value
								Gradient boosting	F1	0.87
								Adaptive boosting	F1	0.90
Nakano et al., 2018		Oral malodour	90	45	45	37	Genus	SVM	Accuracy	0.79
								Deep learning	Accuracy	0.97
Asgari et al., 2018	HMP	Five body sites	1,192	NA	NA	20,589	Various	Random forest	F1	0.89
								SVM	F1	0.85
	Gevers et al., 2014	Crohn's disease	1,359	731	628	9,511	Various	Random forest	F1	0.74
								SVM	F1	0.68

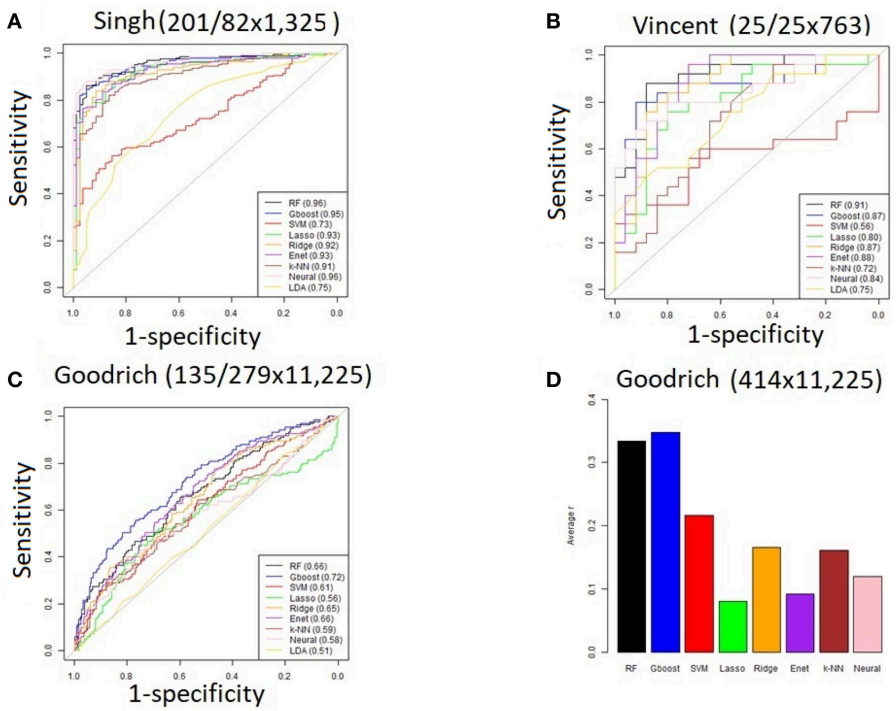


FIGURE 2 | (A–C) ROC curves for each machine learning method using all OTUs. The AUC values are shown in the legend. The size of each dataset (# cases/controls ×# OTUs) is shown in the title. **(D)** Bar graph showing the average Pearson correlation (R) between predicted and actual BMI in the Goodrich dataset, using BMI as a continuous trait.

to the output layer. A bias term is also added in each step, which can be thought of as analogous to the intercept of a linear model. The output layer are predictions based on the data from the input and hidden layers. In most cases, having just one hidden layer with one neuron is reasonable to fit the model.

3.4. Measures of Prediction Accuracy: The AUC and Prediction *R*

For predictions $\hat{y}$ of binary traits, the receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, or a probability of detection. The area under the ROC curve (AUC) is the most common measure of prediction accuracy for binary traits, and ranges from 0.5 (no better than chance) to 1.0 (perfect discrimination). In practice, the empirical AUC can be <0.5, in which case we conclude that the prediction procedure has no value. Note that the AUC is invariant to monotone transformations of $\hat{y}$.

The prediction Pearson correlation (R) between cross-validated predicted and actual y values is a commonly-used standard of accuracy for continuous traits, although many procedures are designed to minimize the mean-squared prediction error $\sum_i (y_i - \hat{y}_i)^2$. $R \leq 0$ corresponds to no predictive value, and $R = 1$ to perfect prediction. We advocate R as a criterion because it is simple and applicable to many prediction procedures. Some prediction procedures may have an offset or proportional bias in prediction that may harm the mean-squared error, even if R is favorable. A *post-hoc* linear rescaling of the

prediction to "fix" any such bias is straightforward, and we find it simplest to directly use R for comparison.

In the real data analyses below, the predicted $\hat{y}$ represent average predictions over all cross-validation rounds, so the AUC and R values were computed directly on the resulting predictions. Importantly, the use of cross-validation provides for each dataset a measure of actual performance of a prediction method, without relying on theoretical considerations, simulations, or restrictive assumptions that may not be applicable with real data.

4. DATA USED FOR COMPARISONS

4.1. A Literature Review

We conducted a literature review of published host-trait microbiome prediction studies that used cross-validation and reported a measure of prediction accuracy. We conducted a literature review of published host-trait microbiome prediction studies that used cross-validation and reported a measure of prediction accuracy. A full table appears in the **Supplement**, including links to each of the 18 studies with 54 reported datasets represented. As different studies used vastly different protocols for OTU generation and preprocessing, for this main paper we focused on the 17 reported datasets that compared at least two competing measures of prediction accuracy. As different studies used vastly different protocols for OTU generation and preprocessing, for this main paper we focused on the 17 reported datasets that compared at least two competing measures of prediction accuracy. All of the datasets were using human hosts, except for Rousk et al. (2010) (where pH in soil samples was the "trait") and Moitinho-Silva et al. (2017), where microbial abundance in sponges was the trait.

4.2. Analyses of Data Using Competing Methods

In addition, we evaluated the supervised learning methods ourselves using datasets from MicrobiomeHD (https://github.com/cduvallet/microbiomeHD), a standardized database of human gut microbiome studies in health and disease. This database includes publicly available 16S rRNA data from published case-control and other studies and their associated patient metadata. The MicrobiomeHD database and original publications for each of these datasets are described in Duvallet et al. (2017). Raw sequencing data for each study was downloaded and processed through a standardized pipeline.

For our analyses, we analyzed four traits (three binary and one continuous) from three datasets with varying sample sizes

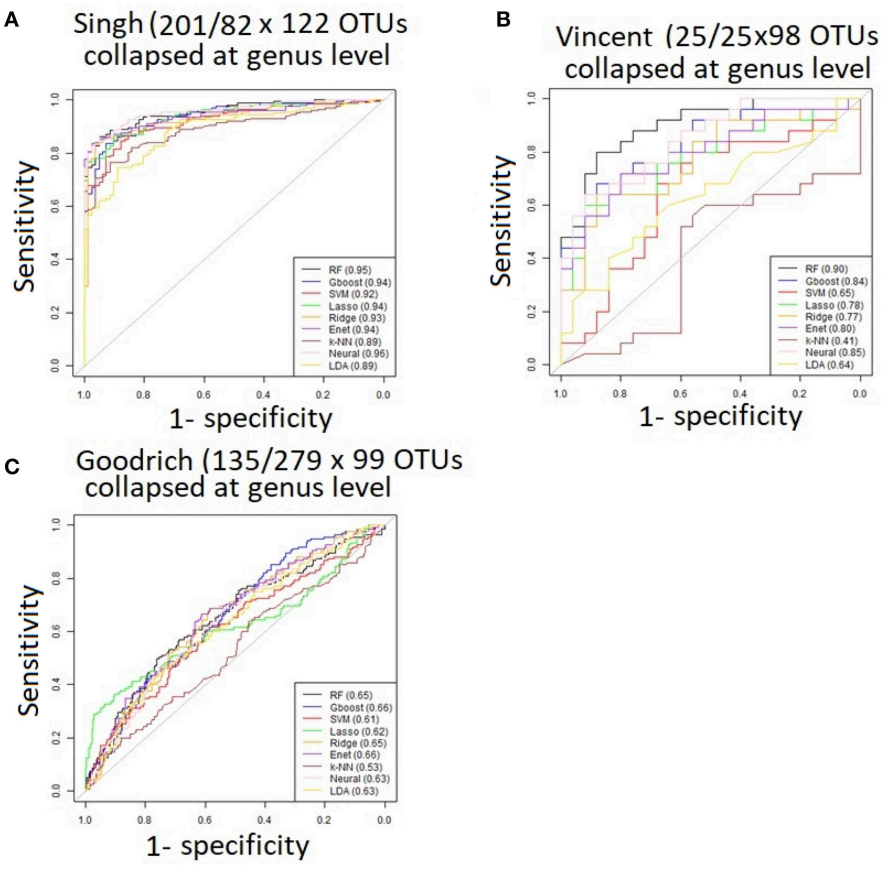


FIGURE 3 | ROC curves after collapsing OTUs to the genus level **(A)** the Singh dataset, **(B)** the Vincent dataset, and **(C)** the Goodrich dataset.

and initial numbers of OTUs: (1) The Singh et al. (2015) data set, containing 201 EDD (enteric diarrheal disease) cases vs. 82 healthy controls with 1,325 OTUs. (2) The Vincent et al. (2013) data set, with 25 CDI (Clostridium difficile infection) cases vs. 25 healthy controls and 763 OTUs. (3a) The Goodrich et al. (2014) dataset, which categorized the hosts into 135 obese cases vs. 279 controls, based on body mass index (BMI), with a total of 11,225 OTUs. In this dataset, individuals came from the TwinsUK population, so we included only one individual from each twin-pair. (3b) The same Goodrich et al. (2014) dataset, but using BMI directly as a continuous phenotype for the same 414 individuals. The microbiome samples for each dataset were obtained from stool, and we analyzed one sample per individual throughout.

Following the filtering recommendations applied by Duvallet et al. (2017), we removed samples with fewer than 100 reads and OTUs with fewer than 10 reads. We also removed OTUs which were present in <1% of samples from the Vincent et al. (2013), Ross et al. (2015), and Singh et al. (2015) datasets, and <5% of samples from the Goodrich et al. (2014) datasets, since it contained many more OTUs. Then we scaled the datasets by calculating the relative abundance of each OTU, dividing its value by the total reads per sample.

In our primary analysis, we tested the relative abundances of the microbiome data at the OTU level. We also ran analyses in which OTUs were collapsed to the genus level by summing their respective relative abundances, discarding any OTUs which were un-annotated at the genus level. Finally, we ran the hierarchical feature engineering (HFE) algorithm introduced by Oudah and Henschel (2018) which results fewer informative features, including individual OTUs and aggregated elements of the taxonomy.

We performed 100 rounds of 5-fold cross-validation for each supervised method, using different random splits for each round. For binary traits, the estimated group probability $\hat{P}(Y = 1|X)$ was used to estimate the group assignment. These estimates were further averaged over the cross-validation rounds. Performance was evaluated using the *AUC*. For continuous traits, the direct estimate $\hat{y}$ was used, averaged over cross-validations, with performance criterion *R*.

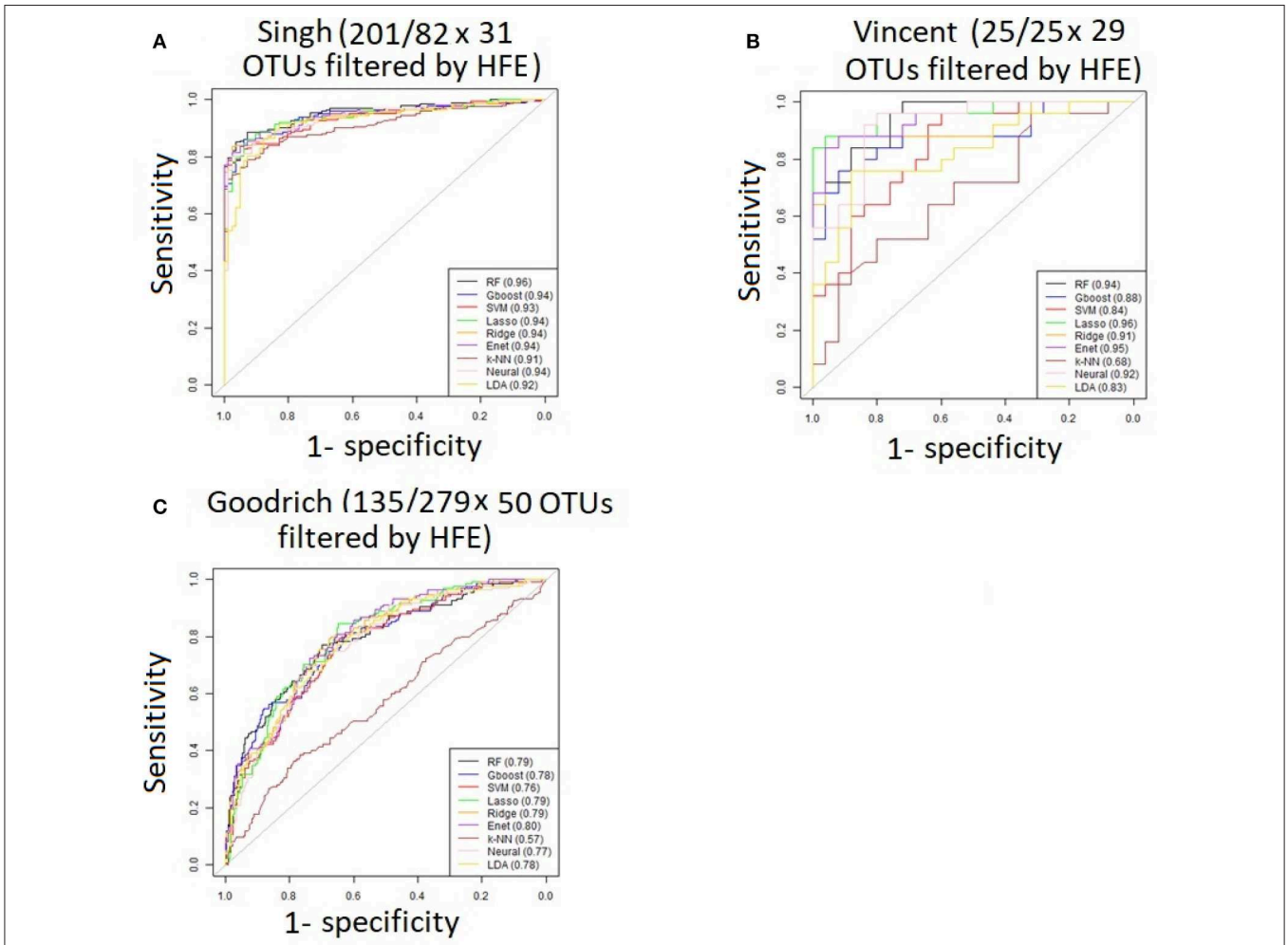


FIGURE 4 | ROC curves after applying the HFE method to select a subset of informative features **(A)** Singh dataset, **(B)** Vincent dataset, **(C)** Goodrich dataset.

R code for the comparisons is available at https://sites.google.com/ncsu.edu/zhouslab/home/software?, and here we list the packages and settings used. Five-fold cross-validation was used throughout, and we additionally checked for plausibility. For example, the out-of-bag accuracy estimates from the random forest procedure were compared to our cross-validated estimates and shown to match closely. All machine learning methods were used for each dataset as applicable (for example, *LDA* was applicable only for the discrete trait datasets). All predictions used probability estimates for the discrete traits. The random forest method used `randomForest` with `ntree=500, mtry=sqrt(ncol(X))`. The gradient boosting (Gboost) decision-tree approach used xgboost, with `nrounds=10` and `objective= "binary:logistic"` for the discrete trait. For the decision tree method, aspects, such as tree depth used default values. The Lasso, Ridge, and Elastic Net approaches used the package and method `glmnet`, with `lambda=seq(0,1,by=0.1)`. The *k*-NN approach used `caret` with $k = 5$ and default (equal) neighbor weighting. The neural net used `neuralnet` with `hidden=1, linear.output=F`. Linear discriminant analysis used the `lda` package with `tol=0`.

5. RESULTS

Table 1 shows the comparative results of 17 datasets analyzed with numerous prediction methods. The results for discrete traits were presented as AUC, accuracy, or balanced accuracy, but in all instances higher values reflect better performance. Although not all methods were represented in each study, some general conclusions can be made. When random forests were applied, they were either the most accurate or competitive [with the exception of Nakano (2018)] (Nakano et al., 2018). Various forms of neural networks often performed well, although there is some question whether the tuning complexity is warranted. An exception is Rousk (2010) as analyzed by Ditzler et al. (2015b), in which some neural networks (perceptions) performed especially well, but the sample size was small $n = 22$. In the datasets analyzed by Ditzler et al. (2015b), the complexity and number of nodes in neural networks showed little consistent relationship to performance. Most of the studies used some form of higher-level OTU aggregation, sometimes as high as the phylum level.

For the three discrete traits, we plotted one ROC curve from each machine learning method (**Figures 2A–C**). The size of each dataset (number of cases/controls X number of OTUs) is shown in the title. Random forest (RTF) and Gradient boosted trees (Gboost) performed well (AUC >0.85) in predicting cases and controls in the Singh and Vincent datasets. Lasso, ridge, elastic net (Enet), *k*-nearest neighbors (k-NN), and Neural Networks (Neural) performed well in the Singh dataset only. Generally, linear SVM and LDA performed less well, and SVM demonstrated close to chance performance in the Vincent dataset.

Summarizing the results after using BMI as a continuous trait in the Goodrich dataset, the bar graph (**Figure 2D**) shows the

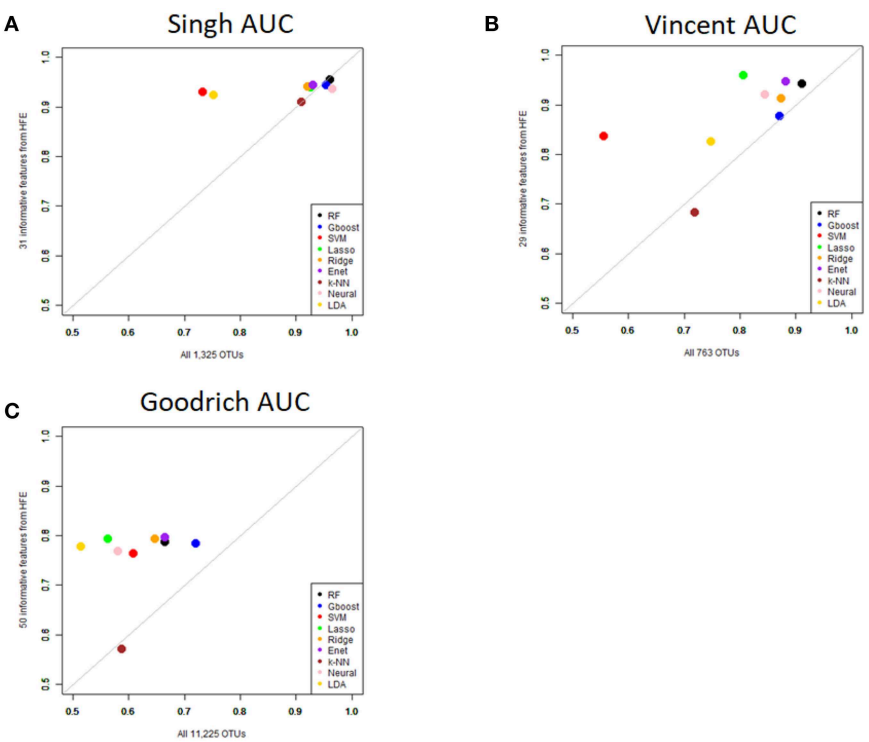


FIGURE 5 | Scatterplot comparing the average AUCs between the full dataset and the HFE subset. **(A)** Singh dataset, **(B)** Vincent dataset, **(C)** Goodrich dataset.

average Pearson correlation between the predicted and actual BMI after 100 iterations of each method. Here again the two decision tree models performed best, although all correlations R were <0.4.

Performance was generally poor for the Goodrich dataset, which also included a large number of OTUs, which presents a challenge in feature selection. We computed the ROC curves for each dataset after collapsing the OTUs to the genus level (**Figure 3**) and after applying the HFE method to select a subset of informative features (**Figure 4**). Then we compared the AUCs between the datasets which used all OTUs and those that used only HFE-informative features (**Figure 5**).

As an overall summary, collapsing to the genus level brought some improvement to the poorer perform prediction methods in the Singh et al. (2015) dataset, and few other broad patterns were apparent. In contrast, the use of cross-validated HFE produced a great improvement in AUC in most instances (**Figure 4**). For the Goodrich et al. (2014) and Singh et al. (2015) datasets, most methods were improved and brought to similar AUC values. For the Vincent dataset, again most prediction methods were improved by HFE feature-reduction, but the results were less uniform. Another pattern that is apparent in the scatterplots, perhaps expected, is that HFE brought diminishing returns for methods that already perform well. The one prediction method that was not improved demonstrably by HFE was k-NN (with $k = 5$).

6. DISCUSSION

We have presented a tutorial overview of the most commonly-used machine learning prediction methods in microbiome host trait prediction. Although a large number of approaches have been used in the literature, some relative simple and clear conclusions can be made. Decision tree methods tended to perform well, and in the published literature similar results were achieved by neural networks and their variants. In our analysis, the HFE OTU feature reduction method brought a substantial performance improvement for nearly all methods. In addition, after such feature reduction most methods performed more similarly. We conclude that this finding accords with the fact that the distinction between sparse and non-sparse methods is less dramatic after feature reduction. We hope that the tutorial, review, and available code are useful to practitioners for host trait prediction.

For more advanced topics, we point the reader to analysis of microbiome time series data, using techniques, such as MDSINE (Bucci et al., 2016), which uses dynamical systems inference to estimate and forecast trajectories of microbiome subpopulations. Other uses of dynamical systems have concentrated mainly on observable phenotypes/experiemental conditions, rather than using microbiome status for prediction (Brooks et al., 2017). In addition, the use of co-measured features, such as metabolites (Franzosa et al., 2019), offers potentially useful information for integrative analyses. As another example of the use of ancillary information, an intriguing approach has also been used to predict biotransformation of specific drugs and xenobiotics by gut bacterial enzymes (Sharma et al., 2017). We also note that our review/tutorial has for clarity placed feature engineering, which may be viewed as a form of statistical regularization, as a separately-handled issue from the penalized prediction modeling. Some modern sparse regression and kernel modeling methods seek additional predictive ability by combining feature regularization and prediction in a single step, e.g., Xiao et al. (2018).

AUTHOR CONTRIBUTIONS

Y-HZ is the leader of this review study. Her contribution includes writing the manuscript, designing the data analysis, summarizing the result, and software management. PG is responsible for the manuscript writing, implementation of analysis, results summary, and code summary.

FUNDING

This work gets support from the NC State Game-changing Research Initiative Program and CFF KNOWLE18XX0.

ACKNOWLEDGMENTS

Thanks to Mr. Chris Smith for the IT support in Bioinformatics Research Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00579/full#supplementary-material

Supplementary Table 1 | Full table of published prediction accuracies.

REFERENCES

Ai, L., Tian, H., Chen, Z., Chen, H., Xu, J., and Fang, J. Y. (2017). Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget* 8, 9546–9556. doi: 10.18632/oncotarget.14488

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc.* 44, 139–177.

Ananthakrishnan, A. N., Luo, C., Yajnik, V., Khalili, H., Garber, J. J., Stevens, B. W., et al. (2017). Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* 21, 603–610. doi: 10.1016/j.chom.2017.04.010

Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics* 34, i32–i42. doi: 10.1093/bioinformatics/bty296

Breiman, L. (2001). Random forests machine learning. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brooks, J. P., Buck, G. A., Chen, G., Diao, L., Edwards, D. J., Fettweis, J. M., et al. (2017). Changes in vaginal community state types reflect major shifts in the microbiome. *Microb. Ecol. Health Dis.* 28:1303265. doi: 10.1080/16512235.2017.1303265

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.* 17:121. doi: 10.1186/s13059-016-0980-6

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. doi: 10.1038/nmeth.3869

Cani, P. D. (2018). Human gut microbiome: hopes, threats and promises. *Gut* 67, 1716–1725. doi: 10.1136/gutjnl-2018-316723

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Chawla, N. V. (2009). "Data mining for imbalanced datasets: an overview," in *Data Mining and Knowledge Discovery Handbook*, eds O. Maimon and L. Rokach (Boston, MA: Springer), 875–886.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297.

Crookston, N. L. and Finley, A. O. (2008). Yaimpute: an r package for knn imputation. *J. Stat. Softw.* 23:16. doi: 10.18637/jss.v023.i10

de Kok, J. B., Roelofs, R. W., Giesendorf, B. A., Pennings, J. L., Waas, E. T., Feuth, T., et al. (2005). Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab. Invest.* 85, 154–159. doi: 10.1038/labinvest.3700208

Ditzler, G., Morrison, J. C., Lan, Y., and Rosen, G. L. (2015a). Fizzy: feature subset selection for metagenomics. *BMC Bioinformatics* 16:358. doi: 10.1186/s12859-015-0793-8

Ditzler, G., Polikar, R., and Rosen, G. (2015b). Multi-layer and recursive neural networks for metagenomic classification. *IEEE Trans. Nanobiosci.* 14, 608–616. doi: 10.1109/TNB.2015.2461219

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

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* 4:293. doi: 10.1038/s41564-018-0306-4

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

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. doi: 10.1016/j.chom.2014.02.005

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. 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

Ho, T. K. (1995). "Random decision forests," in *Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on* (Montreal, QC: IEEE), 278–282.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics* 12, 69–82.

Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat.* 7:e185. doi: 10.1002/sta4.185

Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M., et al. (2016). A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS ONE* 11:e0167370. doi: 10.1371/journal.pone.0167370

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kinross, J. M., Darzi, A. W., and Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome Med.* 3:14. doi: 10.1186/gm228

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, 410–422. doi: 10.1038/s41579-018-0029-9

Knights, D., Costello, E. K., and Knight, R. (2011a). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi: 10.1111/j.1574-6976.2010.00251.x

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011b). Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe* 10, 292–296. doi: 10.1016/j.chom.2011.09.003

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

Lachenbruch, P. A. and Goldstein, M. (1979). Discriminant analysis. *Biometrics* 35, 69–85. doi: 10.2307/2529937

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546. doi: 10.1038/nature12506

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

Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Brief. Bioinformatics* 13, 669–681. doi: 10.1093/bib/bbs054

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. doi: 10.1038/ismej.2011.139

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C. C., Wu, Y.-C., McCormack, G. P., et al. (2017). Predicting the hma-lma status in marine sponges by machine learning. *Front. Microbiol.* 8:752. doi: 10.3389/fmicb.2017.00752

Nakano, Y., Suzuki, N., and Kuwata, F. (2018). Predicting oral malodour based on the microbiota in saliva samples using a deep learning approach. *BMC Oral Health* 18:128. doi: 10.1186/s12903-018-0591-6

Oudah, M. and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 19:227. doi: 10.1186/s12859-018-2205-3

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

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

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rahman, S. F., Olm, M. R., Morowitz, M. J., and Banfield, J. F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *MSystems* 3:e00123-17. doi: 10.1128/mSystems.00123-17

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., Shojaie, A. (2018). Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* 12, 540–566. doi: 10.1214/17-AOAS1102

Reiman, D., Metwally, A., and Dai, Y. (2017). Using convolutional neural networks to explore the microbiome. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2017, 4269–4272. doi: 10.1109/EMBC.2017.8037799

Robinson, C. M. and Pfeiffer, J. K. (2014). Viruses and the microbiota. *Annu. Rev. Virol.* 1, 55–69. doi: 10.1146/annurev-virology-031413-085550

Ross, M. C., Muzny, D. M., McCormick, J. B., Gibbs, R. A., Fisher-Hoch, S. P., and Petrosino, J. F. (2015). 16s Gut community of the cameron county hispanic cohort. *Microbiome* 3:7. doi: 10.1186/s40168-015-0072-y

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555:210. doi: 10.1038/nature25973

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., et al. (2010). Soil bacterial and fungal communities across a ph gradient in an arable soil. *ISME J.* 4:1340. doi: 10.1038/ismej.2010.58

Schmitt, S., Tsai, P., Bell, J., Fromont, J., Ilan, M., Lindquist, N., et al. (2012). Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J.* 6:564. doi: 10.1038/ismej.2011.116

Sharma, A. K., Jaiswal, S. K., Chaudhary, N., and Sharma, V. K. (2017). A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota. *Sci. Rep.* 7:9751. doi: 10.1038/s41598-017-10203-6

Singh, P., Teal, T. K., Marsh, T. L., Tiedje, J. M., Mosci, R., Jernigan, K., et al. (2015). Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome* 3:45. doi: 10.1186/s40168-015-0109-2

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Comput. Biol.* 3:e116. doi: 10.1371/journal.pcbi.0030116

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 58, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Stat. Sci.* 18, 104–117. doi: 10.1214/ss/1056397488

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. doi: 10.1038/nature06244

Vincent, C., Stephens, D. A., Loo, V. G., Edens, T. J., Behr, M. A., Dewar, K., et al. (2013). Reductions in intestinal clostridiales precede the development of nosocomial clostridium difficile infection. *Microbiome* 1:18. doi: 10.1186/2049-2618-1-18

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., et al. (2018). Metagenomics biomarkers selected for prediction of three different diseases in chinese population. *Biomed. Res. Int.* 2018:2936257. doi: 10.1155/2018/2936257

Xiao, J., Chen, L., Yu, Y., Zhang, X., and Chen, J. (2018). A phylogeny-regularized sparse regression model for predictive modeling of microbial community data. *Front. Microbiol.* 9:3112. doi: 10.3389/fmicb.2018.03112

Xu, Z., Malmer, D., Langille, M. G., Way, S. F., and Knight, R. (2014). Which is more important for classifying microbial communities: who's there or what they can do? *ISME J.* 8:2357. doi: 10.1038/ismej.2014.157

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Zhao, S. and Shojaie, A. (2016). A significance test for graph-constrained estimation. *Biometrics* 72, 484–493. doi: 10.1111/biom.12418

Zhou, Y.-H., Marron, J. S., and Wright, F. A. (2018). Computation of ancestry scores with mixed families and unrelated individuals. *Biometrics* 74, 155–164. doi: 10.1111/biom.12708

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest Statement: 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.

The handling editor and reviewer HM declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Genetics

REVIEW
published: 28 August 2019
doi: 10.3389/fgene.2019.00627

Multitable Methods for Microbiome Data Integration

Kris Sankaran[1]* *and Susan P. Holmes*[2]

[1] Mila, Universite de Montréal, Montréal, QC, Canada, [2] Department of Statistics, Stanford University, Stanford, CA, United States

The simultaneous study of multiple measurement types is a frequently encountered problem in practical data analysis. It is especially common in microbiome research, where several sources of data—for example, 16s-rRNA, metagenomic, metabolomic, or transcriptomic data–can be collected on the same physical samples. There has been a proliferation of proposals for analyzing such multitable microbiome data, as is often the case when new data sources become more readily available, facilitating inquiry into new types of scientific questions. However, stepping back from the rush for new methods for multitable analysis in the microbiome literature, it is worthwhile to recognize the broader landscape of multitable methods, as they have been relevant in problem domains ranging across economics, robotics, genomics, chemometrics, and neuroscience. In different contexts, these techniques are called data integration, multi-omic, and multitask methods, for example. Of course, there is no unique optimal algorithm to use across domains—different instances of the multitable problem possess specific structure or variation that are worth incorporating in methodology. Our purpose here is not to develop new algorithms, but rather to 1) distill relevant themes across different analysis approaches and 2) provide concrete workflows for approaching analysis, as a function of ultimate analysis goals and data characteristics (heterogeneity, dimensionality, sparsity). Towards the second goal, we have made code for all analysis and figures available online at https://github.com/krisrs1128/multitable_review.

Keywords: microbiome, data integration, multiomics, dimensionality reduction, heterogeneity

OPEN ACCESS

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

Reviewed by:
Kui Zhang,
Michigan Technological University,
United States
Jing Ma,
Fred Hutchinson Cancer Research Center, United States

***Correspondence:**
Kris Sankaran
kris.sankaran@umontreal.ca

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *14 October 2018*
Accepted: *17 June 2019*
Published: *28 August 2019*

Citation:
Sankaran K and Holmes SP (2019) Multitable Methods for Microbiome Data Integration. Front. Genet. 10:627. doi: 10.3389/fgene.2019.00627

Most methods in statistics expect data to be available as a single table. To a researcher confronted with multiple sources of data, it might therefore seem most natural to either analyze each source separately, one at a time, or else combine all data into a single, unified table. However, neither of these approaches is entirely satisfactory. First, many scientific problems can only be answered by collecting several complementary measurement types. Indeed, the situation is analogous to using many types of sensors to study a single system from many perspectives. Further, while in certain supervised problems, it is enough to predict a single measurement of interest, with other sources collected primarily to provide better features, there are often additional relational components to the analysis: how do different types of measurements co-vary with one another? Here, it is of interest to provide a representation of the data that facilitates comparisons across tables, rather than just comparing each table with a single response of interest. This richer scientific question motivates the development of methods distinct from those used to analyze a single measurement type at a time.

For more concrete motivation, we consider data from the WELL-China study, which is focused on the relationships between various indicators of wellness (Min et al., 2019). In this study,

1,969 individuals[1] underwent clinical examinations, filled out wellness surveys (covering topics such as exercise, sleep, diet, and mental health, for example), and provided stool samples, used for 16s-rRNA sequencing and metabolomic analysis. To date, 16s-rRNA sequencing data are available for 221 of these participants. Evidently, various interesting relational questions can be investigated using this data source.

For the purpose of illustration, we focus on one relatively narrow question that can be addressed using these data: How is the distribution of lean and fat mass across the body related to patterns of microbial abundance? The measurement types most relevant in this analysis are DEXA scans and 16s-rRNA sequencing abundances. DEXA scans use relative X-ray absorption to gauge the amount of lean and fat body mass within a region of the body being scanned. We have access to these lean and fat body mass measurements at several body sites—arms, legs, trunk, etc.—along with related body type variables, like height, age, and android and gynoid fat measurements. In total, there are 36 of these variables. 16s-rRNA sequencing is a technology for gauging the abundance of different bacterial species in the gut by counting the alignments of reads to the 16s-rRNA gene, a component of all bacterial genomes with enough variation to allow discrimination between different individual species. We have counts associated with 2,565 species across 181 genera, though the vast majority are present in low abundances.

This question of the relationship between lean and fat mass distribution (informally, "body type") and the microbiome is motivated by findings that certain taxonomic groups are over- or underrepresented as a function of an individual's body mass index (BMI) (Ley et al., 2005; Ley et al., 2006; Turnbaugh et al., 2009; Ley, 2010). Further, since the distribution of fat is often more related to underlying biological mechanisms than overall body mass (Matsuzawa, 2008), and since this distribution is mediated by specific metabolic pathways, there is reason to suspect that a joint analysis of DEXA and 16s-rRNA microbial abundance data might yield a more complete view of the relationship between the microbiome and body type.

We use this motivating dataset in the examples that follow. Additional numerical examples, for methods only discussed abstractly in this review, are available in the github repository associated with this paper.

CLASSICAL MULTIVARIATE METHODS

Methods from classical multivariate statistics are a mainstay of single-table microbiome data analysis, so it is natural to revisit them before surveying extensions to the multitable setting. Here, we explore a few of the classically studied multitable methods that fit nicely into the modern microbiome data analysis toolbox. We first describe a naive approach based on Principal Components Analysis (PCA)—naive because it lifts a single-table method to the multiple table setting without any special considerations—before studying approaches that directly characterize covariation across several tables: Canonical Correlation Analysis (CCA), Multiple Factor Analysis (MFA), and Principal Component Analysis with Instrumental Variables (PCA-IV).

The earliest multitable method (CCA) was published in 1936, motivated by the problem of relating prices of groups of commodities (Hotelling, 1936). There are two notable aspects of data analysis in this classical paradigm that no longer hold in modern statistics,

- Even when many samples could be collected, there were typically only a few features for each sample, and it was straightforward to study all of them simultaneously. It is now possible to automatically collect a large number of features for each observation (or subject).
- Before electronic computers had been invented, it was important that all statistical quantities be easy to calculate, typically necessitating analytical formulas for parameter estimates. This is no longer an important limitation due to modern computation.

These changes have driven the development of high-dimensional methods and facilitated the adoption of iterative, more computationally intensive approaches.

Nonetheless, it is worth reviewing these original approaches, both to understand the context for many modern techniques and to have an easy starting point for practical data analysis. Indeed, these more established methods tend to be the most readily available through statistical computing packages and can provide a benchmark with which to compare more elaborate, modern methods.

PCA

The simplest approach to dealing with multiple tables is to combine them into one and apply a single-table method, for example, PCA. That is, write

$$X = \left[X^{(1)} \middle| \ldots \middle| X^{(L)} \right] \in \mathbb{R}^{n \times p},$$

where $p = \sum_{l=1}^{L} pl$, and compute the SVD $X = UDV^{\mathrm{T}}$. The K-principal component directions are the first K columns $v_1, \ldots, v_K$, while the associated scores are reweighted rows $d_1 u_1, \ldots, d_K u_K$. We call this method concatenated PCA.

While this does not account for the multitable structure of the data, it does accomplish two goals:

- Through the principal component scores, it provides a visualization of the relationships between samples, based on all features.
- Through the principal component directions, it gives a way of relating features within and across the multiple tables.

However, two drawbacks of this approach are worth noting:

- It does not provide a summary of the relationship between the sets of variables defining the tables—it can only relate pairs of variables.

[1] Though sampling is still ongoing.

- If some tables have many more variables than others, they can dominate the resulting ordination.

These limitations are addressed by CCA and MFA, discussed in sections CCA and MFA, respectively.

We provide one geometric and one statistical motivation for PCA. The geometric motivation is that, if each row x_i of X is viewed as a point in p-dimensional space, then the principal component directions provide the best K-dimensional approximation to the data. The second interpretation is that PCA finds a low-dimensional representation of the x_i such that the resulting points have maximal variance. Qualitatively, this is a desirable property, because it means that the simpler representation preserves most of the variation present in the original data.

PCA is a very widely used technique, and some standard references include Mardia et al. (1980), Friedman et al. (2001), and Pagés (2014). Nonetheless, it is not ideal in the multitable setting.

Example

Figure 1 illustrates this approach on body composition and bacterial abundance data from the WELL-China study. Note that we have subsetted to only women, since men and women have very different body compositions, and we have slightly more data for women. Further, the 16s-rRNA data have been variance stabilized according to the methodology proposed in Anders and Huber (2010) and filtered to only those species that have count ≥5 in at least 7% of samples.

The left panel of **Figure 1** displays the loadings associated with this concatenated PCA approach, where body composition (36 columns) and 16s-rRNA abundances (372 columns) were combined into one dataset (408 columns). Columns associated with bacterial species are displayed as points, shaded by taxonomic family, while columns associated with body composition variables are labeled with text. Note that the fraction of variance explained by each axis is on the order of a few percent—this is to be expected, considering that the baseline proportion would be $\frac{1}{408} \approx 0.25\%$ in the orthogonal case.

Most body composition variables lie close to the vertical axis, in a direction approximately orthogonal to the main direction of variation among species. Columns that are highly correlated—e.g., right (R) and left (L) leg fat mass (FM)—have loadings nearly equal to one another. Among species, the most notable pattern is the concentration of Ruminococcaceae on the right.

To identify relationships between species and body composition variables, it would be of interest to isolate those species with large contributions along the axis defined by linking the center of the variables and the origin. Relatively few such species stand out, though note that there is nothing in this algorithm's objective that would seek covariation across tables directly, so the fact that such associations seem weak with respect to the top two principal components does not mean such relationships do not exist.

We can study individual samples with respect to these loadings, by plotting their projections onto the top two principal components. This is the content of the right panel of **Figure 1**, which displays samples in the same positions, but shaded by android (i.e., abdominal) fat mass. This shading

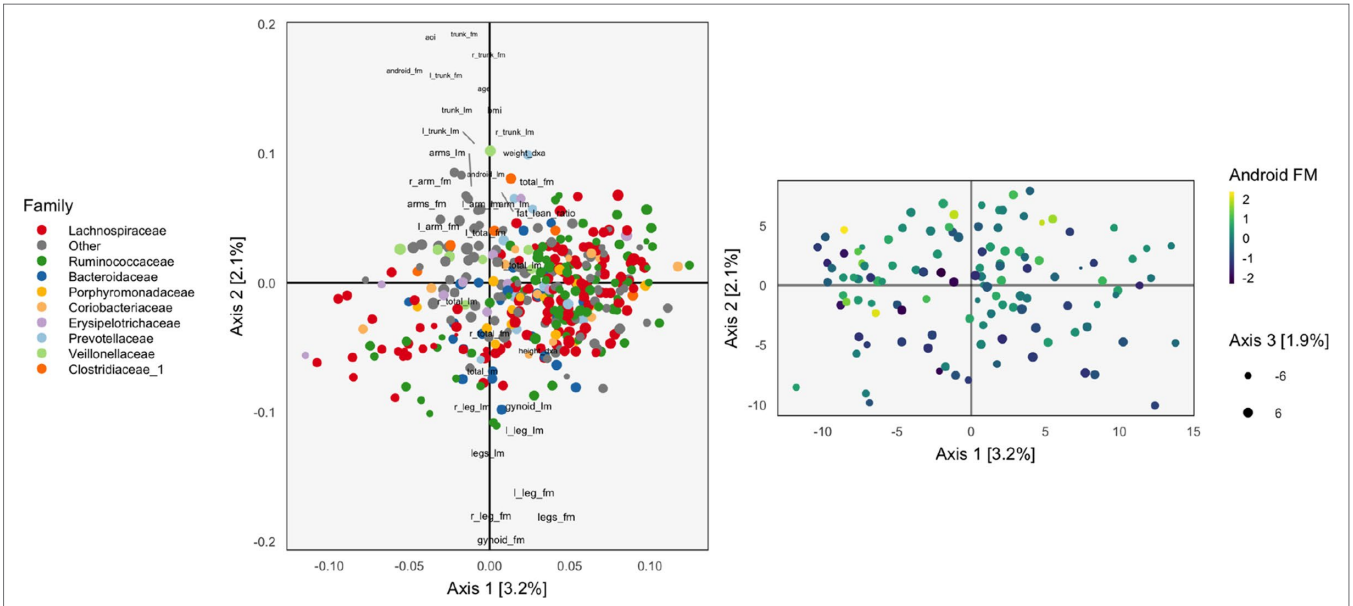


FIGURE 1 | The loadings (left) and scores (right) obtained by applying Principal Components Analysis (PCA) to the combined body composition and microbial abundance data. For the loadings, species are points, and are shaded in by taxonomic family. Body composition variables are plotted as text. The size of points and words measures the contribution of the third PC dimension. For scores, each point corresponds to a sample.

confirms the observations from the loadings directly using observed data. Indeed, the increasing android fat mass among samples in the top of the scores in that panel exactly corresponds to the fact that related variables lie at the top in the left panel.

In this approach, the loadings provide a description of the relationship between variables across datasets. Further, scores summarize variation in samples across multiple datasets. Hence, this heuristic is a natural first step in analyzing multiple table data. However, considering the difficulty in directly interpreting the covariation across datasets, as well as the method's failure to use any sense of covariation in the dimensionality reductions strategy, suggests that this method should not be the last step of an analysis workflow. Nevertheless, we now have a baseline with which to compare the more elaborate methods of subsequent sections.

CCA

CCA is a close relative of PCA, designed to compare sets of features across tables. Like PCA, it provides low-dimensional representations of observations, but it also allows comparisons at the table level. Suppose for now that there are only two tables of interest, $X \in \mathbb{R}^{n\times p_1}$ and $Y \in \mathbb{R}^{n\times p_2}$. Let $\hat{\Sigma}_{XX}, \hat{\Sigma}_{YY}$, and $\hat{\Sigma}_{XY}$ be the associated covariance estimates. Take the SVD, $\hat{\Sigma}_{XX}^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} = \tilde{U}D\tilde{V}^T$. The canonical correlation directions associated with the two tables are $u_k \sum_{XX}^{-\frac{1}{2}} \tilde{u}_k \in \mathbb{R}^{p_1}$ and $v_k = \sum_{YY}^{-\frac{1}{2}} \tilde{v}_k \in \mathbb{R}^{p_2}$. These directions give two sets of low-dimensional representations for each sample, one for each table: $z_k^{(1)} = Xu_k \in \mathbb{R}^n$ and $z_k^{(2)} = Yv_k \in \mathbb{R}^n$. If the two tables are closely related, then the $z_k^{(1)}$ and $z_k^{(2)}$ will be very correlated. The singular values d_k are called the canonical correlation coefficients. Like the eigenvalues in PCA, they characterize the amount of covariation across tables that can be captured by each additional pair of directions.

As with PCA, there are many ways to view this procedure—here we discuss geometric, statistical, and probabilistic interpretations. Unlike the geometric interpretation of PCA, the geometric interpretation for CCA identifies point locations with features, not samples. Specifically, the columns of X and Y are thought of as points in $\mathbb{R}^n$. Consider two subspaces spanning the columns of X and Y, respectively. These subspaces correspond to the linear combinations of features within each table. Place two ellipses on the respective subspaces, centered at the origin and with size and shape depending on the within-table covariances $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$. The first canonical correlation directions are the pair of points, one lying on each ellipse, such that the angle from the origin to those two points is smallest. In this sense, it finds a pair of variance-constrained linear combinations of features within the two tables such that the two combinations appear "close" to one another. The second pair of canonical correlation directions identify a pair of points with a similar interpretation, except they are required to be orthogonal to the first pair, with respect to the inner product induced by the covariances in each table.

For a statistical interpretation, the idea of CCA is to find the low-dimensional representations of the two tables with maximal covariance—this is analogous to the maximum variance interpretation. Formally, rows of the two tables are imagined to be i.i.d. draws from $\mathbb{P}^{XY}$, which has marginals $\mathbb{P}^X$ and $\mathbb{P}^Y$. Consider arbitrary linear combinations $z_i^{(1)}(u) = u^T x_i$ and $z_i^{(2)}(v) = v^T y_i$ of samples from the two tables. The first pair of CCA directions u_i^* and v_i^* are chosen to optimize

$$\begin{aligned} &\underset{u\in\mathbb{R}^{p_1}, v\in\mathbb{R}^{p_2}}{\text{maximize}} \quad \mathrm{Cov}_{\mathbb{P}^{XY}}\left[z_i^{(1)}(u), z_i^{(2)}(v)\right] \\ &\text{subject to } \mathrm{Var}_{\mathbb{P}^X}\left(z_i^{(1)}(u)\right) = 1 \\ &\qquad\qquad \mathrm{Var}_{\mathbb{P}^Y}\left(z_i^{(2)}(v)\right) = 1 \end{aligned} \tag{1}$$

To produce subsequent directions, the same optimization is performed, but with the additional constraint that the directions must be orthogonal to all the previous directions identified for that table. Of course, in actual applications, we estimate these covariances and variances empirically.

This perspective makes it easy to derive the algorithm given at the start of this section. The empirical version of the optimization problem (1) is

$$\begin{aligned} &\underset{u\in\mathbb{R}^{p_1}, v\in\mathbb{R}^{p_2}}{\text{maximize}} \quad u^T \hat{\Sigma}_{XY} v \\ &\text{subject to } u^T \hat{\Sigma}_{XX} u = 1 \\ &\qquad\qquad v^T \hat{\Sigma}_{YY} v = 1. \end{aligned} \tag{2}$$

Consider the transformed data, $\tilde{u} = \hat{\Sigma}_{XX}^{\frac{1}{2}} u$ and $\tilde{v} = \hat{\Sigma}_{YY}^{\frac{1}{2}} v$. The optimization can be now be expressed as

$$\begin{aligned} &\underset{\tilde{u}\in\mathbb{R}^{p_1}, \tilde{v}\in\mathbb{R}^{p_2}}{\text{maximize}} \quad \tilde{u}^T \hat{\Sigma}_{XX}^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} \tilde{v} \\ &\text{such that } ||\tilde{u}||_2^2 = 1 \\ &\qquad\qquad ||\tilde{v}||_2^2 = 1. \end{aligned} \tag{3}$$

The optimal $\tilde{u}_1$ and $\tilde{v}_1$ for this problem are well known—they are exactly the first left and right eigenvectors of $\hat{\Sigma}_{XX}^{\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} = \tilde{U}D\tilde{V}^T$, respectively.

A probabilistic interpretation of this procedure views it as estimating the factors in an implicit latent variable model. In particular, (Bach and Jordan, 2005) supposes that x_i and y_i are drawn i.i.d. from the model,

$$\begin{aligned} \xi_i := (\xi_i^S, \xi_i^x, \xi_i^y) &\sim \mathcal{N}(0, Id) \\ x_i \,|\, \xi_i &\sim \mathcal{N}(\mu_x + W_X \xi_i^S + B_X \xi_i^x, I_d) \\ y_i \,|\, \xi_i &\sim \mathcal{N}(\mu_Y + W_Y \xi_i^S + B_Y \xi_i^y, I_d) \end{aligned}$$

That is, each sample is associated with a d-dimensional latent variable ξ_i, drawn from a spherical normal prior. A few of the coordinates of these latent variables, ξ_i^s, contribute to shared structure, through W_X and W_Y. The remaining coordinates model table-specific structure, through B_X and B_Y. It can be shown that the posterior expectations of the latent ξ_i^s given the observed tables must lie on the subspace defined by the CCA directions.

Example

We next apply CCA to the WELL-China body composition and microbiome data, with particular interest in how the results compare with those of section Example. We provide analogous loadings and scores plots in **Figure 2**. However, note that the data are not quite the same between the two analysis—we have filtered down to species passing a filter, which reduces the number of species to 66, from 2,565. This very aggressive filtering is necessary because CCA requires estimation of covariances matrices, and Σ_{XX}, Σ_{XY}, and Σ_{YY}, which is impossible for $p > n$ and highly unstable when p is a large fraction of n. Besides this stronger filtering, all preprocessing steps remain the same as in section Example.

The left panel of **Figure 2** provides the analog of CCA loadings. To be precise, let $X \in \mathbb{R}^{102\times36}$ be the matrix of body composition measurements and $Y \in \mathbb{R}^{102\text{x}66}$ be the variance-stabilized microbial abundances. As before, write $u_k \in \mathbb{R}^{36}$, $v_k \in \mathbb{R}^{66}$ for the k^{th} canonical correlation directions. Text labels from column j of the body composition variables are displayed at position $(u_{j1}, u_{j2})_{j=1}^{36}$ and shaded points for the j^{th} species at position $(v_{j1}, v_{j2})_{j=1}^{66}$.

As in the concatenated PCA, we find that the groups of variables occupy separate spaces. Our interpretation is that sequences further to the left are correlated with the body variables further to the left, which are all in some way variants of body mass. Note that age is negatively correlated with total fat mass, which is why it appears on the opposite end. Among the abundant species that remain, there is limited clustering according to taxonomic group, though the Bacteroideceae and Ruminoccocus do appear restricted to the bottom right and left, respectively.

In the right panel of **Figure 2**, we plot the corresponding scores. Note that in CCA, there are two sets of scores for each k, the Xu_k and Yv_k. Indeed, the CCA objective finds directions that maximize the correlation between these scores. We use a different color legend for the two panels, each of which represents one set of scores. The legend for scores from species abundances are colored by family, while those for the body composition associates samples with android fat mass. The pairs of scores for each individual sample are drawn with small links. Since most links are relatively short, linear combinations of the two tables could be found that optimized the objective—indeed, the top two canonical correlations are 0.968 and 0.957. However, some caution is necessary here, and a more honest evaluation would be based on scores obtained by projecting new samples onto the original CCA directions. This is especially important in this nearly high-dimensional setting, where covariance estimation may be unreliable.

Aside from the fact that samples appear as pairs, interpretation proceeds as in a PCA scores plot, as in **Figure 1**. The association between these variables and the sample positions is not as strong as when performing PCA on the combined table. This is to be expected, however, as PCA maximizes variance without any thought to covariance, and the body composition table alone has a large portion of its variance related to android fat mass.

Co-Inertia Analysis

Co-inertia Analysis (CoIA) emerged in ecology to facilitate analysis of variation in species abundance as a function of environmental conditions (Dolédec and Chessel, 1994). It can be viewed as a slight modification of CCA. Again, we seek sets of orthonormal directions $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ such that the associated projections Xu_k and Yv_k explain most of the covariation between the tables. Unlike CCA, CoIA finds its first directions by maximizing the covariance—not the correlation—between scores,

$$\underset{u\in\mathbb{R}^{p_1},v\in\mathbb{R}^{p_2}}{\text{maximize}} \quad u^T X^T Y v$$

$$\textit{such that } \|u\| = 1$$

$$\|v\| = 1,$$

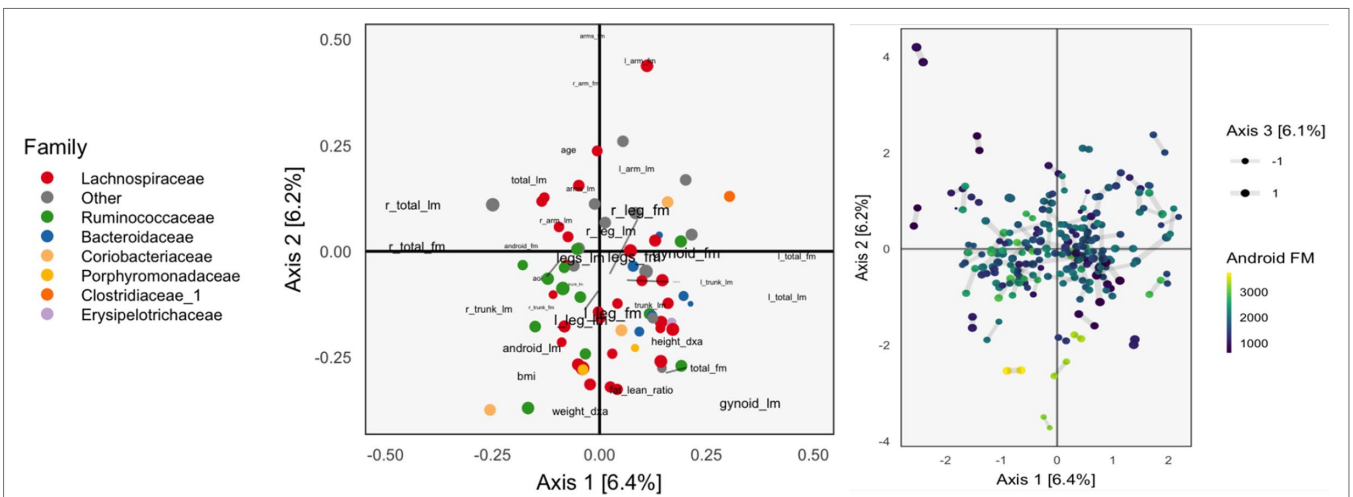


FIGURE 2 | The Canonical Correlation Analysis (CCA) analog of the PCA biplot in **Figure 1**, obtained by applying CCA to the combined body composition and microbial abundance data. Since each sample is associated with a pair of scores, one from each table, we use a different symbol to represent the scores: two points joined by an edge, where each point gives the score from one of the tables. Aside from this exception, the PCA biplot interpretation still applies. The higher the CCA objective, the shorter the links between pairs. The first two CCA dimensions suggest smooth variation across samples, according to amount of android fat mass.

with subsequent directions found by the same optimization, after adding the constraint that they are orthogonal to the previously derived directions.

The only difference with the objective in equation (2) is that norm constraint is imposed on u and v directly, rather than their transformations $\sum_{XX}^{\frac{1}{2}} u$ and $\sum_{YY}^{\frac{1}{2}} v$. It is in this sense that the CCA objective maximizes the correlation between scores, while CoIA maximizes the covariance.

The solution $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ can be obtained as the first K left and right eigenvectors from the SVD of X^TY, as opposed to the first K generalized eigenvectors, as in CCA. The proof of this fact is almost identical to the derivation in section CCA, for CCA.

Example

We apply CoIA to the same data as used in section Example, as CoIA also needs to estimate the covariance between tables, which is difficult when the number of species is large. We find that the associated scores are quite different from those found using CCA. Compare **Figure 3**, which shades samples by android fat mass with **Figure 2** for CCA. The scores for CoIA are not so closely aligned across tables, but they exhibit a clearer gradient across android fat mass. We find that the scores are not nearly as closely aligned as they are for CCA, but that they are more strongly associated with variation in android fat mass, as in the concatenated PCA result of **Figure 1**. It is not clear whether this phenomenon—the CoIA scores being more similar to those from PCA than CCA—holds in general, or what it is about the change in inner products between CoIA and CCA that is responsible for this difference.

MFA

MFA gives an alternative approach to producing scores and relating features across multiple tables (Pagés, 2014). It can be understood as a refined version of the concatenated PCA described in section PCA that reweights tables in a way that prevents any one table from dominating the resulting ordination. Specifically, MFA is a concatenated PCA on the matrix

$$X := \left[\frac{1}{\lambda_1(X^{(1)})} X^{(1)} | \ldots | \frac{1}{\lambda_1(X^{(L)})} X^{(L)} \right],$$

which reweights each table $X^{(k)}$ by its largest eigenvalue, $\lambda(X(k))$. This procedure is the multitable analog of the common practice of standardizing variables before performing PCA.

The resulting MFA directions and scores can be interpreted in the same way as those from PCA—the MFA directions still specify the relationship between measured features, and the position of each sample's projection describes the relative value of each feature for that sample. Moreover, MFA gives a way of comparing entire tables to each other, called a "canonical analysis" (Pagés et al., 2004). A K-dimensional representation of the l^{th} group is given by

$$\left[\mathcal{L}\left(z_1, X^{(l)}\right), \ldots, \mathcal{L}\left(z_K, X^{(l)}\right) \right],$$

where $z_k = d_k u_k \in \mathbb{R}^n$ is the k^{th} column of principal component scores and

$$\mathcal{L}\left(z_k, X^{(l)}\right) = \frac{\lambda_k(X)}{\lambda_1(X^{(l)})} \text{tr}\left(X^{(l)} X^{(l)T} z_k z_k^T\right) = \frac{\lambda_k(X)}{\lambda_1(X^{(l)})} \| X^{(l)T} zk \|_2^2$$

is a measure of aggregate similarity between the coordinates in the l^{th} table and the k^{th} column of scores. According to this definition, if the samples, as represented by the l^{th} table, have high correlation with the k^{th} dimension of scores, then the canonical analysis displays positions the l^{th} table far in the k^{th} direction. Plotting these table-level coordinates helps resolve which tables measure similar underlying variation.

PCA-IV

PCA-IV adapts the dimensionality reduction ideas of PCA to the multivariate regression setting (Rao, 1964). It can also be

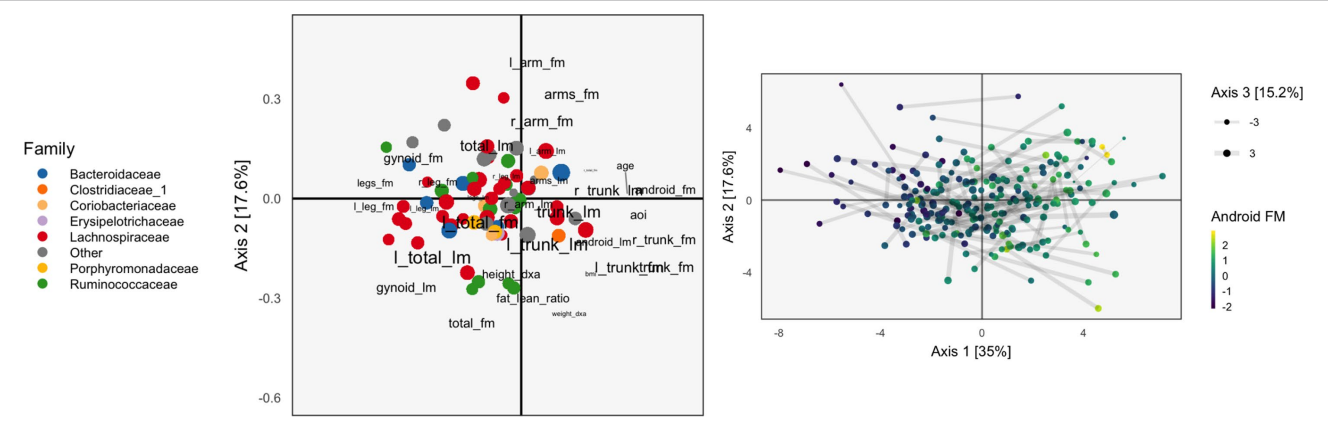


FIGURE 3 | The Co-inertia Analysis (CoIA) analog of the PCA and CCA biplots in **Figures 1** and **2**. There seems to be a clearer gradient across android fat mass variables, though the scores are not so well aligned, since the links are somewhat longer.

viewed as a version of PCA that chooses a dimension reduction of X based on its ability to predict Y. In this sense, it anticipates methods like Partial Least Squares, Canonical Correspondence Analysis, the Curds & Whey procedure, and the Graph-Fused Lasso, which are described in sections Partial Least Squares, CCpnA, Curds & Whey, and Graph-Fused Lasso.

Formally, suppose we are predicting $y_i \in \mathbb{R}^{p_1}$ from $x_i \in \mathbb{R}^{p_2}$. Since p_2 may be large, it might be useful to work with a lower-dimensional representation $z_i = V^T x_i \in \mathbb{R}^K$, which is potentially more interpretable but still as (or more) predictive of y_i. As in PCA, we require that V be orthonormal.

The criterion that PCA-IV uses to identify the loadings V and scores Z mirrors the maximum variance criterion for PCA. Instead of choosing V to maximize the variance of the z_i, we choose it to minimize the residual covariance of y_i given z_i. That is, suppose that y_1 and x_1 are jointly normal with mean 0 and covariance

$$\mathrm{Var}_{\mathbb{P}}\begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}.$$

If $z_i = VTx_i$, then the joint covariance of y_i and z_i is

$$\mathrm{Var}_{\mathbb{P}}\begin{pmatrix} y_i \\ z_i \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX}V \\ V^T\Sigma_{XY} & V^T\Sigma_{XX}V \end{pmatrix},$$

so the residual covariance of y_1 given z_1 is

$$\Sigma_{YY} - \Sigma_{YX}V\left(V^T\Sigma_{XX}V\right)^{-1}V^T\Sigma_{XY}. \tag{4}$$

Rao (Rao, 1964) uses the trace to measure the "size" of this matrix. The true population covariances are unknown to us, so we replace them by their empirical estimates. The formal optimization for PCA-IV then becomes

$$\underset{V\in\mathbb{R}^{p_2\times K}\text{orthonormal}}{\text{minimize}} \ \mathrm{tr}\left(\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX}V(V^T\hat{\Sigma}_{XX}V)^{-1}V^T\hat{\Sigma}_{XY}\right) \tag{5}$$

The optimal V are the top K generalized eigenvectors of $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}$ with respect to $\hat{\Sigma}_{XX}$, that is, the orthonormal set of (v_k) satisfying

$$\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}V = \left(\lambda_1\hat{\Sigma}_{XX}v_1 \,|\ldots|\, \lambda_k\hat{\Sigma}_{XX}v_k\right) = \hat{\Sigma}_{XX}V\Lambda,$$

where $\Lambda = \mathrm{diag}\,(\lambda_k) \in \mathbb{R}^{K\times K}$. A derivation for why this choice is optimal is provided in section *Derivation Details for PCA-IV*.

For a geometric interpretation of PCA-IV, view each column y_j in Y and x_j in X as a point in $\mathbb{R}^n$. Assuming X and Y are full rank, the collections (y_j) and (x_j) span p_1- and p_2-dimensional subspaces. A set of independent regressions of y_j on X projects each individual y_j onto the span of the (x_j), and the squared residuals are the distance to this subspace. The PCA-IV procedure is an attempt to find a further K-dimensional subspace within the span of the (x_j) such that the residuals of the regressions from y_j onto this further subspace is not much worse. This is displayed in **Figure 4**.

Example

Continuing our WELL-China case study, we now illustrate results from PCA-IV. The idea of scores and loadings in this context requires some clarification. By PCA-IV scores, we mean the coordinates of projections z_i of samples onto the subspace defined by V, and by loadings, we mean the correlation between columns[2] of X and Y with the PCA-IV axes defining V.

The scores and loadings are given in **Figure 5**. Interpretation of the species loadings is simple, since species seem well separated by taxa. Interpretation of the body composition variables is less clear—pairs of variables that would be expected to be near to one another are not, in many cases. Indeed, leg fat mass (leg_fm) and left leg fat mass (l_leg_fm) should have a small angle between one another, but they do not. It is possible that by approximating the covariation across tables, the quality of within-table approximations deteriorates.

We find that the scores, displayed in figures, are similar to those that found by the concatenated PCA of section PCA. One possible explanation for this behavior is that the PCA-IV-generalized SVD of X is similar to an ordinary PCA of X, and that in the concatenated PCA of $(Y\ X)$, the fact that X has many more columns than Y means that the result is similar to a PCA on X alone.

Partial Triadic Analysis

Partial Triadic Analysis (PTA) gives an approach for working with multitable data when each table has the same dimension, $p1 = p2$ (Kroonenberg, 2008; Thioulouse, 2011). Specifically, it gives a way of analyzing data of the form $(X_{\cdot\cdot l})_{l=1}^{L}$, where each $X_{\cdot\cdot l} \in \mathbb{R}^{n\times p}$. This is called a data cube because it can also be written as a three-dimensional array $X \in \mathbb{R}^{n\times p\times L}$. We denote the j^{th} feature measured on the i^{th} sample in the l^{th} table by x_{iji}, and the slices over fixed i, j, and l by $X_{i\cdot\cdot}$, $X_{\cdot j\cdot}$, and $X_{\cdot\cdot l}$. This type of data arises frequently in longitudinal data analysis, where the same features are collected for the same samples over a series of L times. However, the actual ordering of the L tables is not ever used by this method: if we scrambled the time ordering for L tables, the algorithm's result would not change.

The main idea in PTA is to divide the analysis into two steps:

- Combine the L tables into a single compromise or consensus table.
- Apply any standard single-table method, e.g., PCA, on the compromise table.

A naive approach to constructing the compromise table would be to average each entry across the L tables. Instead, PTA upweights tables that are more similar to the average table, as these are considered more representative. Formally, the compromise is defined as $X_c = \Sigma_{l=1}^{L}\alpha_l X_{\cdot\cdot l} = X\alpha \in \mathbb{R}^{n\times p}$, where α (constrained to norm one) is chosen to maximize $\sum_{l=1}^{L}\alpha_l\left\langle \bar{X}, X_{\cdot\cdot l}\right\rangle$,

[2]Geometrically, the angle between original columns and the subspace, in the sense of **Figure 4**.

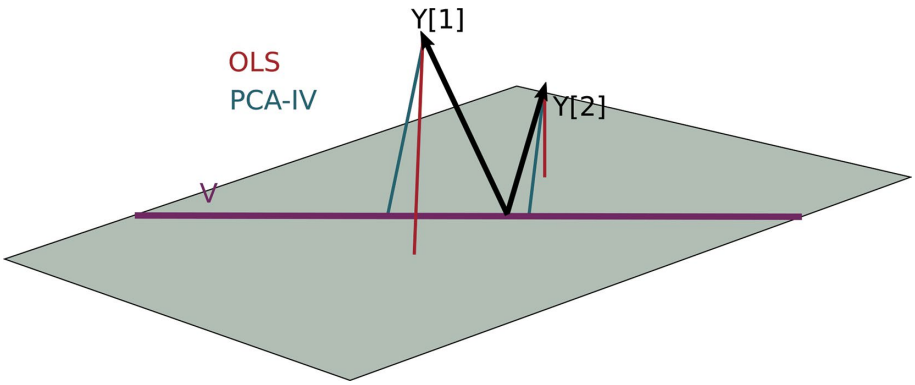


FIGURE 4 | A geometric view of Principal Component Analysis with Instrumental Variables (PCA-IV). The columns of the response Y are views as n-dimensional vectors. The gray plane is the span of X. Multivariate OLS simply projects the columns of Y onto the plane, while PCA-IV searches for a further subspace V on which to project all responses.

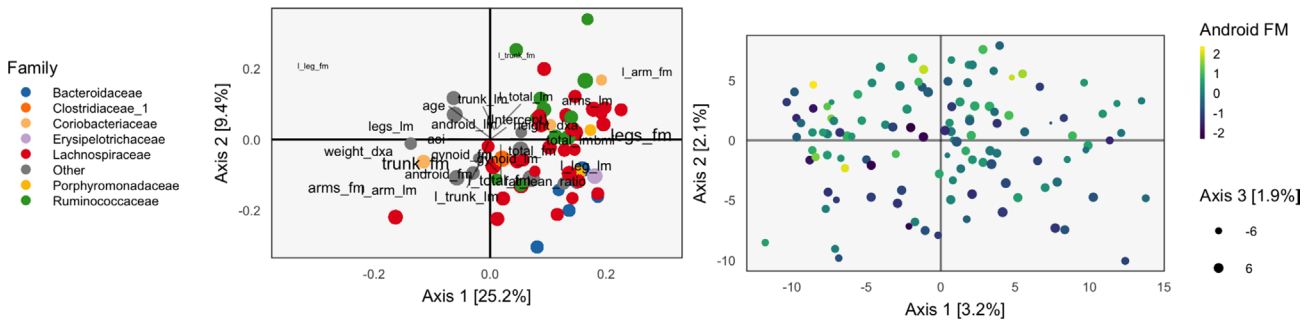


FIGURE 5 | The PCA-IV biplot can be interpreted like biplots from previous methods, for example, **Figure 1**. Some of the relationships between variables seem less intuitive than those observed previously.

a weighted average of inner-products[3] between each of the L tables and the naive-average table, $\bar{X} = \frac{1}{L}\sum_{l=1}^{L} X_{\cdot\cdot l}$.

The optimal α can be derived using Lagrange multipliers (see *Derivation of PTA α*) and leads to the compromise table,

$$X_c = \sum_{l=1}^{L} \frac{\left\langle \bar{X}, X_{\cdot\cdot l} \right\rangle}{\sqrt{\sum_{l'=1}^{L} \left\langle \bar{X}, X_{\cdot\cdot l'} \right\rangle^2}} X_{\cdot\cdot l}.$$

We can try to interpret the compromise matrix geometrically. Suppose the $X_{\cdot\cdot l}$ define an orthonormal basis, so that $\left\langle X^l, X^{l'} \right\rangle = \mathbb{I}(l = l')$. Then, we can write the compromise table as

$$X_c = \sqrt{L}\sum_{l=1}^{L} \left\langle \bar{X}, X_{\cdot\cdot l} \right\rangle X_{\cdot\cdot l} = \sqrt{L}\bar{X},$$

a scaled version of the mean.

If, however, the tables are not orthonormal, then we place more weight on directions that are correlated. For example, if $X^{(1)} = X^{(2)}$, but the rest of the tables are orthogonal to each other and to these first two tables, then the compromise double counts the direction $X^{(1)}$. Therefore, compared to the naive average $\bar{X}$, X_c upweights more highly represented tables.

Statico and Costatis

In the multivariate ecology literature, it is common to have a pair of data cubes, giving species abundances and environmental variables over time, respectively. We write these as $Y \in \mathbb{R}^{n \times p_1 \times L}$ and $Y \in \mathbb{R}^{n \times p_2 \times L}$. Costatis and Statico are two approaches for analyzing such data (Thioulouse, 2011). They are easiest to understand as divide-and-conquer approaches, where the general problem of analyzing a pair of data cubes is divided into two steps, one designed for analyzing individual cubes, and another for studying covariation across tables. In Statico, the covariation problem is dealt with first, then followed by a data cube analysis, while in Costatis, that order is reversed.

Specifically, in Statico, an empirical cross-covariance matrix is constructed at each time point, $Z^l = \frac{1}{n_l} Y_{\cdot l}^T X_{\cdot l}$. For example, this is the correlation between the environmental variables and species counts at a specific time point l. The L matrices Z^l are then

[3] We are using $\langle A, B \rangle = \mathrm{tr}(A^T B)$.

input into a PTA, yielding a compromise table Z_c that can then be studied with PCA.

Alternatively, in Costatis, a compromise table is constructed for each of the data cubes Y and X, using PTA. Call these Y_c and X_c. These are now simply two matrices, each with n rows, and they can be analyzed by any two-table dimensionality reduction method, for example, CoIA.

Hence, we see that the only difference between these methods is the order in which CoIA and PTA are applied. Indeed, this is reflected in the names of the methods: Statis is an abbreviation for a PTA, and Statico performs a CoIA before a Statis while Costatis does the reverse.

MODERN MULTIVARIATE METHODS

Compared to classical approaches, modern multivariate methods are typically designed for more high-dimensional, heterogeneous settings. The two methods reviewed in this section are examples of this trend: Partial Least Squares (PLS) is well-suited for finding predictors in the presence of high-dimensional response matrices, while Canonical Correspondence Analysis (CCpnA) was designed to facilitate joint analysis of heterogeneous continuous and count data necessary. Unlike traditional statistical methods, neither approach is explicitly model-based, and both are iterative, requiring more extensive computation than earlier techniques.

Partial Least Squares

PLS sequentially derives a set of mutually orthogonal features $(z_k)_{k=1}^K$ that characterizes the relationship between two tables, Y and X (Wold, 1985). To obtain the first PLS direction, z_1, compute the first left singular vector u_1 of the cross-covariance matrix between the two tables, $\hat{\Sigma}_{YX} = \frac{1}{n} Y^T X$. Then, for each of the p_2 columns of X, compute the univariate (i.e., partial) regression coefficient $\hat{\varphi}_j = \frac{1}{\| x_{\cdot j} \|_2^2} x_{\cdot j}^T u_1$, for $j = 1, \ldots, p_1$. The first PLS direction is defined as $z_1 = \sum_{j=1}^{p_2} \hat{\varphi}_j x_{\cdot j}$ a weighted average of $x_{\cdot j}$ according to their partial correlation with u_1. To generate subsequent directions z_k, orthogonalize both Y and X with respect to the current directions $z_1, \ldots z_{k-1}$, and repeat the process.

This procedure is appealing because, like PCA, it reduces a potentially high-dimensional matrix X with many correlated columns into a smaller set of orthogonal directions. Moreover, it achieves this reduction in a way that accounts for correlation with columns in Y: columns of X that are uncorrelated with Y will have no contribution to the PLS directions, even if they account for a large proportion of variation in X.

We have stated the procedure in the form it was originally proposed, but this algorithmic description is difficult to understand geometrically or probabilistically. However, interpretational aids have since been developed. Frank and Friedman (1993) and Stone and Brooks (1990) studied the case where $p_1 = 1$, so y is a single column vector. By assuming that the rows of y and X are drawn i.i.d. from distribution $\mathbb{P}^{YX}$, with marginals $\mathbb{P}^Y$ and $\mathbb{P}^X$, they found that the k^{th} PLS direction z_k is the z that solves the optimization

$$\begin{aligned} &\underset{z}{\text{maximize}}\ \text{Corr}_{\mathbb{P}^{YX}}\left[x_i^T z_k, y_i\right] \text{Var}_{\mathbb{P}^X}(z^T x_i) \\ &\text{such that } z^T X^T X z_j = 0 \text{ for all } j \le k-1 \\ &\quad \| z \|_2 = 1. \end{aligned} \tag{6}$$

If the covariance term is omitted, the optimization is identical to the maximum variance problem that gives the principal component directions based on X. This formulation makes precise the idea that PLS is a version of principal components that accounts for correlation with Y.

An alternative interpretation, due to (Gustafsson, 2001), is that PLS fits a particular latent variable model. Suppose $\xi_i = (\xi_i^s, \xi_i^X)$ are drawn i.i.d. from a $K_1 + K_2 = K$ dimensional spherical normal. PLS assumes the observed tables Y and X have rows drawn i.i.d. from

$$\begin{aligned} y_i \mid \xi_i &\sim \mathcal{N}(\mu_Y + W_Y \xi_i^s, \sigma^2 I_{p_1}) \\ x_i \mid \xi_i &\sim \mathcal{N}(\mu_X + W_X \xi_i^s + B_X \xi_i^X, \sigma^2 I_{p_2}). \end{aligned}$$

That is, each table is the sum of two components, one that is a table-specific linear combination of a shared latent variable, and another that is an arbitrary linear combination of a table-specific latent variable. The shared feature ξ^s is the object of interest, and is what PLS implicitly estimates.

Sparse Partial Least Squares

PLS suffers from two of the same problems as PCA:

- It can be unstable in high-dimensional settings, since it requires estimation of covariances, and isn't well defined when $p > n$.
- PLS directions are linear combinations of all features in x_i, which can be difficult to interpret when there are many features.

Different regularized, sparse modifications of PCA have been proposed to remedy these issues in the PCA context (Jolliffe et al., 2003; Zou et al., 2006; Witten et al., 2009). For PLS, similar analysis leads to sparse PLS (Lê Cao et al., 2008; Chun and Kele, 2010), and we briefly review this method here.

Directly regularizing the multiresponse version of the PLS optimization (6) leads to the problem

$$\begin{aligned} &\underset{z_k}{\text{maximize}} \sum_{j=1}^{p_1} \text{Cov}_{\mathbb{P}^{YX}}\left[x_i^T z_k, y_{ij}\right] \\ &\text{such that } z^T \boldsymbol{x}^T x z_j = 0 \text{ for all } j \le k-1 \\ &\quad \| z_k \|_2 = 1 \\ &\quad \| z_k \|_1 \le \lambda, \end{aligned}$$

which can be applied to real data by replacing the objective with its sample version, $z_k^T M z_k$, where $M = X^T Y Y^T X$. This version

of the problem falls into the Penalized Matrix Decomposition framework of Witten et al. (2009), reviewed in the section penalized matrix decomposition.

However, Chun and Kele (2010) argue that this formulation does not lead to "sparse enough" solutions. Instead, they adapt the SPCA approach of Zou et al. (2006) to PLS. The resulting objective identifies two sets of directions, a set (a_k) that maximizes the PLS-defining covariance and another, (z_k), that approximates the first set by a sparser alternative. Formally,

$$\begin{aligned} \underset{z_k, a_k}{\text{maximize}} & -\kappa \| a_k \|_M^2 + (1-\kappa) \| z_k - a_k \|_M^2 \\ \text{such that } & \| a_k \|_2^2 = 1 \\ & \| z_k \|_1 \leq \lambda_1 \\ & \| z_k \|_2 \leq \lambda_2, \end{aligned} \tag{7}$$

where we have defined $\| x \|_M = \sqrt{x^T M x}$ and κ, λ_1, and λ_2 are tuning parameters. The first term in the objective is the PLS-defining covariance, the second ensures that the solutions z_k and a_k are similar, and the norm constraints induce sparsity and stability on z_k. Note that while this objective is not convex, for fixed a_k, it is an elastic-net regression, while for fixed z_k, it is a type of eigenvalue problem.

Example

Next we apply the sparse partial least squares (SPLS) implementation of Chung et al. (2012) to the WELL-China body composition data. We use the body composition variables as the response Y and the microbiome community composition as X. In this direction, a well-fitting model would allow the microbiome community measurements X to serve as a proxy for the variables in Y, in case those data were not easily accessible. To an extent, however, this choice of directionality is arbitrary—regressing abundances on body composition variables would also be sensible—and reflects the basic limitations of using an asymmetric method to study a symmetric problem.

We subset to female subjects and filter species, keeping only those species with a count of at least 5 in at least 7% of samples. This leaves 372 species over 119 participants. All species abundances are variance-stabilized using the approach of Anders and Huber (2010). We cross-validate with five folds, searching through a grid over $K \in \{4,\ldots,8\}$ and $\lambda_1 \in \{0, 0.05,\ldots, 0.7\}$. This grid is used to prevent the model from regularizing to the point that there is no information to visualize. For example, if we set $K = 1$, every row of **Figure 6** would look identical. The predictive accuracy is poor, which is unsurprising considering the spike at 0 in the abundances histogram—the held out error is ≈ 1.29, after having scaled and centered the body composition variables.

Figure 6 displays fitted coefficients relating body composition variables with species abundances. By fitted coefficients, we mean we display $\hat{B} = ZQ^T$, where Z are the SPLS directions and a multiresponse linear regression model is used. Specifically, $Y = XB + E = XZQ^T + E$ where X is a matrix with rows x_i, Y is a matrix with columns y_j, and Z is a matrix with columns z_k.

Positive associations tend to occur across all responses simultaneously, while negative associations can be unique to either lean or fat mass. Most taxonomic families seem to have slightly more negative than positive associations, with the possible exception of Porphyromonodaceae.

To interpret these coefficients in the raw data, we can visualize individual species with strong associations to body composition. Specifically, we study associations with the android and gynoid fat mass variables. In the left panel of **Figure 7**, we display the abundances X for species against android fat mass, respectively. The species are chosen according to whether the two-dimensional coefficient across android and gynoid fat mass has large norm[4]. The main associations that are visible are those between the body composition and species presence or absence. That is, there don't seem to be any cases where a body composition feature varies smoothly as a species becomes more or less abundant. Instead, SPLS has identified species whose samples have lower or higher android or gynoid fat mass, depending on whether that species is present or absent.

CCpnA

CCpnA is a method, originally developed in ecology, useful for joint analysis of count and continuous data. The canonical application has a site-by-species count matrix $Y \in \mathbb{R}^{n \times p_1}$ and an environmental features matrix $X \in \mathbb{R}^{n \times p_2}$, for example, historical rainfall and temperature measurements. In the WELL context, Y would be the samples by community abundance matrix, while X would contain the body composition measurements.

The scientific goal might be to identify species that are more abundant in sites with more rainfall or higher temperature. If these environmental variables were uncorrelated, it would be enough to fit a separate regression to each. This, however, is rarely the case, motivating the development for CCpnA.

Translating to the language of the WELL-study, individual samples can be thought of sites, and the supplemental data—that is, the body composition variables—are analogous to environmental variables.

CCpnA produces low-dimensional representations of both the rows and columns of Y (the samples and species), along with latent subspaces on which these representations are defined. Algorithmically, CCpnA first constructs the following matrices, where 1_r denotes a column vector of r ones,

1. An overall frequency matrix,

$$F = \frac{1}{n_{\cdot\cdot}^Y} Y,$$

where $n_{\cdot\cdot}^Y$ is the sum of all counts in matrix Y.
2. A diagonal matrix of row (site) proportions,

$$D_r = \text{diag}(F_{p_1}^1) \in \mathbb{R}^{n \times n}.$$

[4] Specifically, $\left\| \begin{pmatrix} \beta_{android} \\ \beta_{gynoid} \end{pmatrix} \right\|_2 > 0.065.$

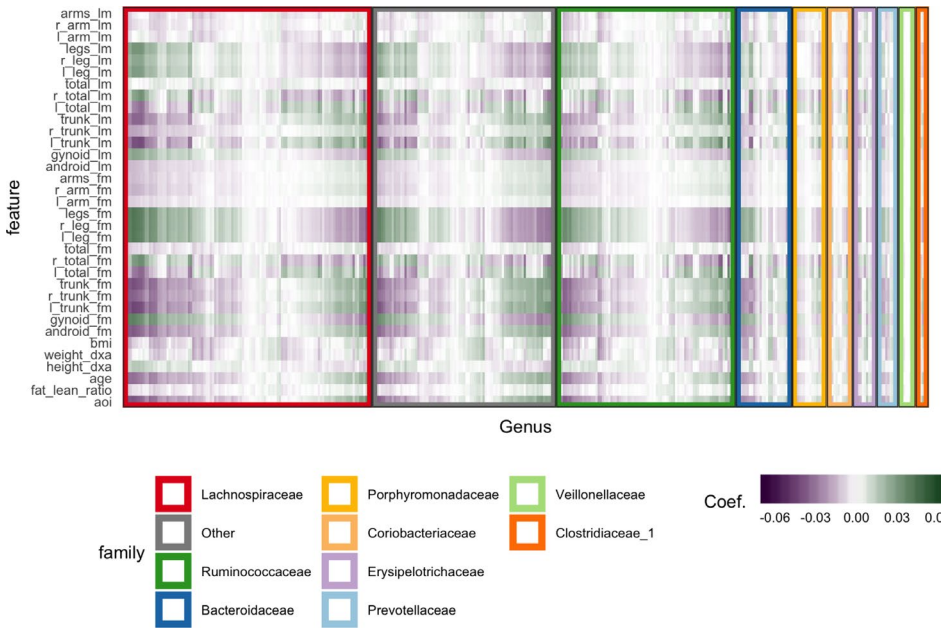


FIGURE 6 | Coefficients learned by SPLS. Each row is a response dimension, which is a body composition variable. Each column is associated with a species. The shading within each cell corresponds to the SPLS coefficient for that species–response pair. Green and purple cells are positive and negative coefficients, respectively. Species are grouped first according to their taxonomic family, marked by grouping panel colors, and then by a hierarchical clustering on coefficient values.

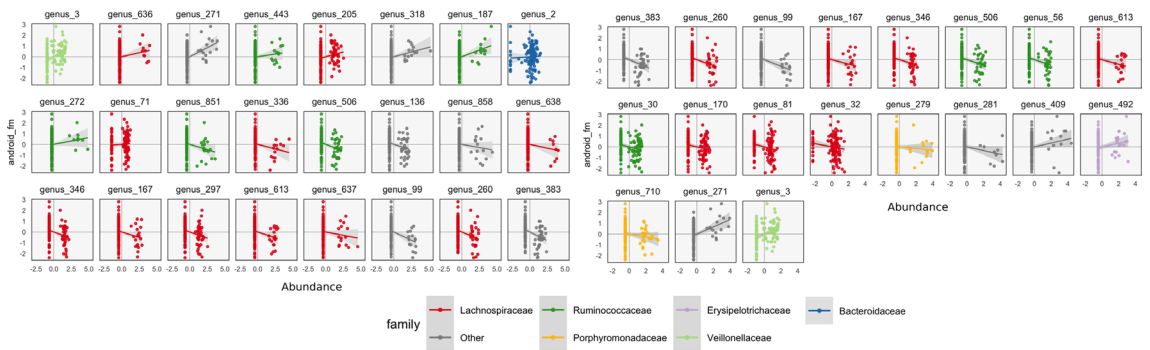


FIGURE 7 | A more focused view of the species with high loadings according to SPLS (left) and sparse CCA (right). Each panel corresponds to a species. Points are shaded according to each species' taxonomic family. The *x*-axis within panels corresponds to variance-stabilized species abundance, while the *y*-axis gives android fat mass. A linear smooth is provided to summarize the direction of associations. Panels are arranged according to the size of that species' absolute SPLS coefficient value or loading onto the first sparse CCA axis. The presence of certain species seems to correspond to increased or decreased levels of android fat mass.

3. A diagonal matrix of column (species) proportions,

$$D_c = \mathrm{diag}(F^T 1_n) \in \mathbb{R}^{p_1 \times p_1}.$$

4. A projection onto the columns of the supplemental matrix X, reweighting samples according to their species counts,

$$P_X = D_r^{-1/2} X \left(X^T D_r X\right)^{-1} X^T D_r^{-1/2} \in \mathbb{R}^{n \times n},$$

With this notation, compute an SVD,

$$D_r^{-\frac{1}{2}} \left(F - F 1_{p_1} 1_{p_1}^T F\right) D_C^{-\frac{1}{2}} P_X = USV^T,$$

and define row and column scores Z and Q by

$$Z = D_r^{-\frac{1}{2}} US$$

$$Q = D_c^{-\frac{1}{2}} V^T S.$$

There are several ways to interpret this procedure. CCpnA was originally proposed as the solution to a fixed-point

iteration called reciprocal averaging (Ter Braak, 1986). Later, Greenacre (1984) and Greenacre and Hastie (1987), provided a geometric view and Zhu et al. (2005) gave an exact probabilistic interpretation.

The intuition for the reciprocal averaging procedure is simple: the scores for different samples should be a weighted average of the species scores, with larger weights for the species that are more common at those sites. Similarly, species scores can be defined according to a weighted average of sample scores. That is,

$$z_i \propto \frac{1}{f_{i\cdot}} \sum_{j=1}^{P1} f_{ij} q_{ij}$$

$$q_i \propto \frac{1}{f_{\cdot j}} \sum_{i=1}^{n} f_{ij} z_{ij},$$

or, in matrix form,

$$Z \propto \text{diag}\,(F1p_1)^{-1} F Q^T$$

$$Q \propto \text{diag}\,(F^T 1_n)^{-1} Z.$$

This formulation suggests an algorithm for finding Z and Q—arbitrarily initialize one and iterate these calculations until convergence.

As is, this is not yet the setup that yields CCpnA—it does not use information in the supplemental table X. To recover CCpnA, a projection step needs to be inserted before the calculation of row scores,

1. Arbitrarily initialize Z.
2. While not converged,
   a. Solve $Q' \propto \text{diag}\,(F^T 1_n)^{-1} F^T Z$.
   b. Project $Q = P_X Q'$.
   c. Solve $Z \propto \text{diag}\,(Z1_{p_1})^{-1} F Q^T$.

The fixed point of this iteration is the previously described CCpnA solution.

A second interpretation is due to Zhu et al. (2005). Suppose first that we are only interested in a one-dimensional score for rows and columns. Let α be a latent gradient, for example, between warm-dry and cold-wet sites, or low and high android-fat mass samples. For each of the p_1 species, define a normal density over the supplemental variables, $f_j(x_i) = \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$. The mode of this density represents the preferred environment for species j. Next, project these densities onto the gradient, giving a univariate $f_j^{\alpha}(z_i) = \mathcal{N}(z_i \mid \alpha^T \mu_j, \alpha^T \Sigma_j \alpha)$ for each species. The z_i represent the scores for species i along the gradient α.

The generative model views species–sample pairs one at a time. For each pair involving sample i and species j, draw a score according to $f_j^{\alpha}(z_i)$. Hence, each site i draws species according to a p_1-class linear discriminant (LDA) model.

To use this idea to compute scores, we need to estimate the gradient α, which is also of interest in its own right. This is done by supposing equal covariances across species, $\Sigma_j = \Sigma$ for all j, and finding the $\hat{\alpha}$ maximizing the between vs. total variance across species,

$$\frac{\alpha^T \Sigma_B \alpha}{\alpha^T \Sigma \alpha},$$

where

$$\Sigma_B = \sum_{j=1}^{p_1} f_{\cdot j} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T$$

is a between-species covariance matrix. Estimating $\hat{\alpha}$ in this way and writing $z_i = \hat{\alpha}^T x_i$ gives the original site scores from CCpnA.

We have omitted a detailed numerical example of this method in this review, but note that codes for applying this method are available in the github repository associated with this review.

Penalized Matrix Decomposition

In high-dimensional settings, sparsity is a desirable property, for both qualitative interpretability and statistical stability. A regression model using only a few features is easier to understand than one involving a linear combination of all possible features. Further, regularized models typically outperform their unregularized counterparts in terms of both predictive accuracy and inferential power (Buhlmann and Van De Geer, 2011). In fact, it is impossible to fit an unregularized linear regression when the number of features is greater than the number of samples.

The Penalized Matrix Decomposition (PMD) is a general approach to adapting the regularization machinery developed around regression to the multivariate analysis setting (Witten et al., 2009). The CCA and MultiCCA instances of PMD have been particularly well-studied (Witten et al., 2009; Witten et al., 2013).

The general setup is as follows. Suppose we want a one-dimensional representation of the samples (rows) in $X \in \mathbb{R}^{n \times p}$. Recall that the first k-eigenvectors recovered by PCA span a subspace that minimizes the ℓ^2-distance from the original data to their projections onto that subspace. In particular, when $k = 1$, the associated PCA coordinates $u \in \mathbb{R}^n$ and eigenvector v are the optimal values in the problem

$$\underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p, d \in \mathbb{R}}{\text{minimize}} \; \| X - duv^T \|_2^2$$

$$\text{subject to } \| u \|_2^2 = \| v \|_2^2 = 1.$$

The PMD generalizes this formulation of rank-one PCA to enforce additional structure on u and v. The PMD solutions u and v are defined as the optimizers of

$$\begin{aligned} &\underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p, d \in \mathbb{R}}{\text{minimize}} \; \| X - duv^T \|_2^2 \\ &\text{subject to } \| u \|_2^2 = \| v \|_2^2 = 1. \\ &\qquad \text{Pen}_u(u) \le \mu_1 \\ &\qquad \text{Pen}_v(v) \le \mu_2 \end{aligned} \tag{8}$$

where Pen_u and Pen_v are arbitrary constraints u on and v.

To choose the regularization parameters μ_1 and μ_2, Witten et al. (2009) applied cross-validation to the reconstruction errors after holding out random entries in X. To obtain a sequence of scores $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ for $K > 1$, define u_k and v_k as the optimizers of the problem (equation 8) on the residual: $X^k := X^{k-1} - d_{k-1}u_{k-1}v_{k-1}^T$ where $d_k = u_k^T X^k v_k$ and $X^1 = X$.

This view can be specialized to develop regularized versions of a number of multivariate analysis problems. We consider applications to the CCA and MultiCCA problems. Recalling that $\| A \|_F^2 = \text{tr}(A^T A)$ along with the linearity and the cyclic properties of the trace, the objective in equation (8) can be rewritten, using $\equiv$ to mean equality up to terms constant in u and v,

$$\begin{aligned} \| X - duv^T \|_F^2 &= \text{tr}\left(\left(X - duv^T\right)^T \left(X - duv^T\right)\right) \\ &\equiv -2d\,\text{tr}\left(X^T uv^T\right) + d^2 tr\left(uv^T uv^T\right) \\ &\equiv -2dv^T X^T u + d^2, \end{aligned}$$

where for the last equivalence we used that $v^T v = u^T u = 1$.

From this expression, and by partially minimizing out $d = v^T X^T u$, we see that the PMD solutions u and v in equation (8) can be found as the optimizers of

$$\begin{aligned} &\underset{u\in\mathbb{R}^n, v\in\mathbb{R}^p}{\text{maximize}} u^T X^T v \\ &\text{subject to } \|u\|_2^2 = \|v\|_2^2 = 1 \\ &\qquad \text{Pen}_u(u) \le \mu_1 \\ &\qquad \text{Pen}_v(v) \le \mu_2 \end{aligned}$$

Notice that, as long as the penalties are convex in u and v, the optimization is biconvex, so a local maximum can be found by alternately maximizing over u and v.

From this form, we can derive a sparsity-inducing version of CCA. Recall the maximal-covariance interpretation of CCA,

$$\begin{aligned} &\underset{u\in\mathbb{R}^{p_1}, v\in\mathbb{R}^{p_2}}{\text{maximize}} \; u^T \hat{\Sigma}_{XY} v \\ &\text{subject to } u^T \hat{\Sigma}_{XX} u = v^T \hat{\Sigma}_{YY} v = 1 \end{aligned}$$

Witten et al. (2009) argue for diagonalized CCA, in which the variance constraints are replaced by unit norm constraints, and sparsity-inducing ℓ^1 constraints are added,

$$\begin{aligned} &\underset{u\in\mathbb{R}^{p_1}, u\in\mathbb{R}^{p_2}}{\text{maximize}} u^T \hat{\Sigma}_{XY} v \\ &\text{subject to } \|u\|_2^2 = \|v\|_2^2 = 1 \\ &\qquad \|u\|_1 \le \mu_1 \\ &\qquad \|u\|_1 \le \mu_2 \end{aligned}$$

which is exactly of the form of equation (9) where $X = \hat{\Sigma}_{XY}$.

Multiple CCA can also be described in this framework, by replacing the objective with the sum over all pairwise covariances, $\sum_{l,l'=1}^L c_1^{(l)T} X^{(l)T} X(l') c_1^{(l')}$, and introducing constraints for each of the $c_1^{(l)}$.

Example

We apply the PMD formulation of sparse CCA to the WELL-China data. As before, we k-over-A filter the microbiome data, requiring species to have counts of at least 5 in at least 7% of samples. Further, we first variance-stabilize, center, and scale these species abundances. For the regularization parameters, we set $\mu_1 = 0.7$ for the body composition data and $\mu_2 = 0.3$ for the species count data. The reasoning behind the relative values of these two tuning parameters is that sparsity in species loadings is more important than sparsity across body composition variables, because the microbiome data are more high-dimensional. The choice of the tuning parameters' overall magnitude was guided by the overall number of factors that we wanted to retain.

We only compute the first three PMD directions, and the associated correlations between scores are $(d_1, d_2, d_3) = (0.700, 0.435, 0.632)$. Note that the correlation can increase in subsequent directions, since directions are computed iteratively and cannot be defined and sorted all at once.

The learned loadings and scores are displayed in **Figure 8**. The x-axis in the loadings differentiates between high android and gynoid fat mass. The y-axes in the loadings reflect a gradient between overall right and left body mass. The size of points corresponds to the third PMD direction, and it seems to highlight high BMI, ratio of fat to lean mass, and overall weight. We interpret species based on their positions relative to these body composition variables, as in an ordinary biplot. For example, genus 492, located in the center-top, seems to be more common among people with higher android and lower gynoid fat mass.

The associated scores are displayed in the right panel, shaded according to android fat mass. The gradient between android and gynoid fat mass suggested by the loadings is clearly visible from this display. The length of links reflects the correlation between sets of scores. They are somewhat longer in the sparse CCA compared to the ordinary CCA on a subset of species, but this is likely a consequence of regularization and overfitting on the part of ordinary CCA.

We can follow up these displays by focusing on species that seemed related to the CCA axes. In the right panel of **Figure 7**, we isolate species with loa dings a distance of at least 0.15 from the origin. These are the same ones that are labeled by text in **Figure 8**. We can see associations between abundance and android fat mass, as suggested by the loadings. Generally, there is a difference between android fat mass among people with and without particular species—there is no smooth function between the quantity of a species android fat mass, even in these cases where an association exists. Further, no individual taxonomic group seems to dominate the set of associated species.

Multitable Mixed-Membership

In section CCA, a latent variable interpretation of CCA was provided as an alternative to the standard covariance maximization perspective. Since likelihood-based methods are easily adapted to different data types, it is natural to consider versions of CCA designed

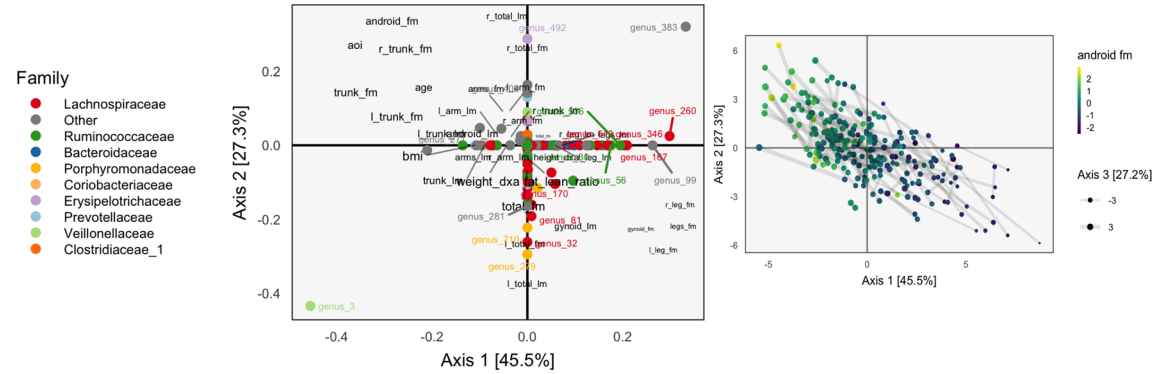


FIGURE 8 | A sparse CCA biplot, for variables with at least one nonzero coordinate. In the loadings (left), each point corresponds to a species, and is shaded in by tax onomic family. Species with loadings far from the origin are also annotated with their names. Black text are loadings for body composition variables. The size of points and text reflects the contribution of the third CCA dimension. Many loadings have at least one dimension that is exactly zero, due to ℓ^1-regularization. For the sample scores (right), each point is a sample, positioned at their coordinates with respect to the first two learned sparse CCA directions. Points are shaded according to android fat mass, and their sizes are set according to the third sparse CCA direction's contribution. Evidently, the first two directions reflect a gradient across android fat mass, suggesting that this is a substantial contributor to covariation across microbiome and body composition tables.

for non-Gaussian data, using section CCA as a starting point. We are particularly interested in data with the same structure as the WELL-China body composition and microbiome data, namely, two table data where one table is continuous with Gaussian marginals and correlated columns and the other is a high-dimensional collection of counts, where many entries are exactly zero.

As before, define a set of shared scores $\xi_i^s \in \mathbb{R}^K$, and two sets of within-table scores $\xi_i^X \in \mathbb{R}^{L_1}$ and $\xi_i^Y \in \mathbb{R}^{L_2}$. As before, we model the body composition variables using essentially a Gaussian factor analysis model, $y_i \,|\, \xi_i^X, \xi_i^Y \sim \mathcal{N}(B^y \xi_i^s + W^y \xi_i^y, \sigma^2 I_{p_2})$ with a spherical Gaussian prior ξ_i^X, ξ_i^Y on. For the counts matrix, we might consider a few different approaches:

- *Bayesian Exponential Family PCA* (Mohamed et al., 2009): By requiring low-rank structure on the natural parameters of an exponential family model, we could naturally model high-dimensional count data, using a Poisson or multinomial likelihood, for example.
- *Nonnegative Matrix Factorization* (Lee and Seung, 2001): A variant of the exponential family approach is to model the counts matrix as a Poisson likelihood over a low-rank product of Gamma random matrices.
- *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003): We can model the observed samples as Dirichlet mixtures of a few underlying "topics," which are themselves drawn from a Dirichlet prior.

Here, we focus on the LDA approach, though we suspect that the other two approaches are potentially interesting as well. Formally, this model supposes that counts are drawn according to

$$x_i \,|\, (\theta_k) \sim \text{Mult}\left(x_i | N_i \sum_{k=1}^{K} \theta_{ik} \beta_k \right)$$
$$\theta_i \sim \text{Dir}\,(\alpha)$$
$$\beta_k \sim \text{Dir}\,(\gamma),$$

where $N_i = \sum_{j=1}^{p_1} x_{ij}$ is the total count in sample i. This has the flavor of a factor analysis where $(\theta_{ik})_{k=1}^K$ are scores for the i^{th} sample and (β_k) are K underlying topics.

The only complexity with using an LDA model of X together with a Gaussian factor analysis on Y is that the shared scores ξ_i^s typically have different priors—a Dirichlet for LDA and a spherical Gaussian for factor analysis. In any formulation of probabilistic CCA that uses both models, this must be reconciled. One approach is to continue to place Dirichlet priors on all the scores, ξ_i^s, ξ_i^x, and ξ_i^y. While the model for the Gaussian data is no longer exactly traditional factor analysis, it has a similar interpretation. Alternatively, we could use a spherical Gaussian prior on all scores and then recover probability vectors by applying the softmax function, $[\mathcal{S}(v)]_k = \dfrac{\exp(v_k)}{\sum k' \exp(v_{k'})}$,

$$x_i | \xi_i^s, \xi_i^x \sim \text{Mult}\left(x_i | N_i, \mathcal{S}\left(B^X \xi_i^s + W^X \xi_i^x \right) \right)$$
$$\xi_i^s \sim \mathcal{N}(\xi_i^s | \mathbf{0}, \tau^2).$$

It is this second model that we use in our experiments below.

Example

We illustrate this multitable mixed-membership approach on the WELL-China data. We choose $K = 2$ for the number of shared topics and $L_1 = L_2 = 3$ for the number of unshared topics per table. We initialize scores and loadings using results from the PMD formulation of sparse CCA. While the use of shared ξ_i^s and unshared (ξ_i^x, ξ_i^y) scores gives more flexibility in modeling, it also leads to additional complexity in interpretation—there are both more scores and more loadings that need to be visualized.

Consider the loadings W^X and W^Y, provided in the left panel of **Figure 9** and bottom three rows of **Figure 10**. Note that there is no notion of variance explained by different axes in this case.

The loadings W^X of **Figure 9** summarize table-specific variation in bacterial abundances. Invariance under rotation and reflection

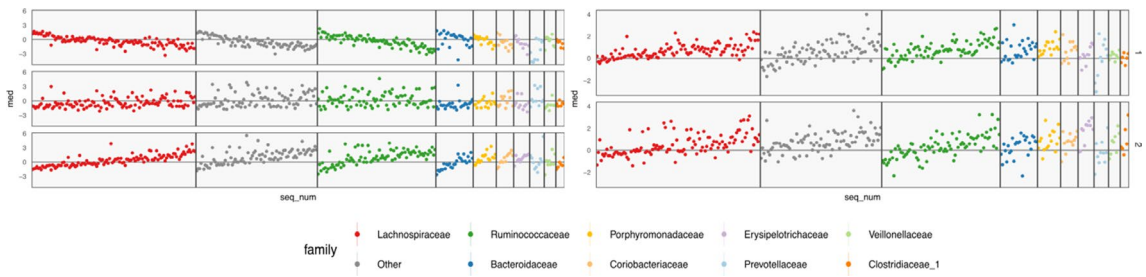


FIGURE 9 | Table-specific (left) and cross-table (right) loadings for different species. Each row is a loading dimension, columns are features (species in this case), and intervals summarize posterior samples for the associated loading parameter, W^X_{jk} for table-specific loadings, and B^X_{jk} for cross-table loadings. Species are sorted from most to least abundant, within each taxonomic family. Caution must be exercised when interpreting these loadings, as loadings are invariant under rotations and reflections.

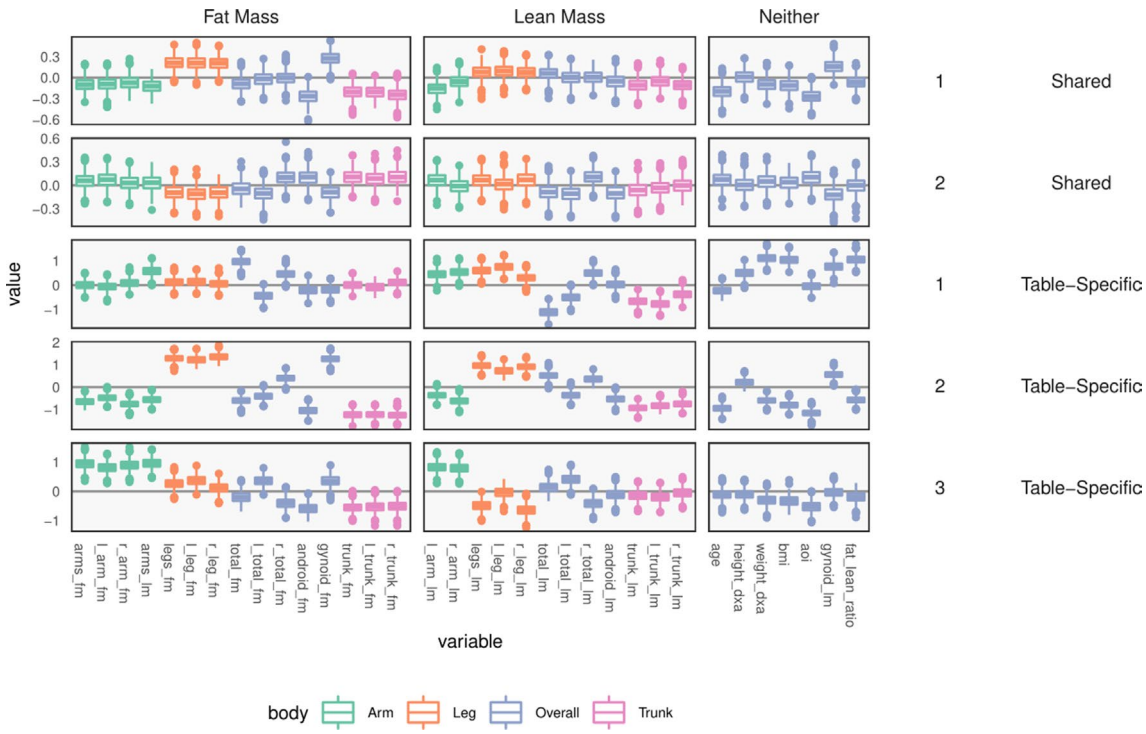


FIGURE 10 | Table-specific and shared loadings, for the body composition variables, corresponding to the parameters W^Y_{jk} and B^Y_{jk} As in **Figure 9**, each row is one loading dimension, columns are features, and boxplots summarize posterior samples for the associated loading parameters. Colors distinguish between parts of the body. We note that loadings learn specific contrasts between types of fat mass and parts of the body.

complicates interpretation of these estimates. If we flip the sign of all the loadings axes, then the more abundant species have larger loadings, so the direction of different trends is irrelevant. The main distinction between the first and second loadings is the rate of decay in frequencies, especially among Lachnospiraceae and Ruminococcaceae. For example, topic 1 seems to include species from these taxonomic families that are not very abundant. The main characteristic of the third loading is that it has higher values for Porphyromonadaceae, so samples with high weight on this loading have decreased levels of these taxa.

Next, consider within-table body composition loadings, given in the bottom three rows of **Figure 10**, which suggests that the first and third axes of W^Y capture variation between overall and android vs. gynoid fat mass. The first axis has high loadings for weight, BMI, and total fat mass, and the third contrasts areas with high android and high gynoid fat mass. The second axis distinguishes between right and left total lean and fat mass variation, while the third axis captures difference between mass in the trunk versus arms and legs.

These summaries could have been obtained by analyzing each table separately. Covariation between the two tables is captured by the shared scores ξ^s_i and loadings B^X, B^Y. The shared body composition loadings are given in the top two rows of **Figure 10**. These loadings again differentiate between android and gynoid fat mass, learning contrasts between body mass in arms and legs,

for example, though the effects are less pronounced than in the table-specific loadings.

The shared bacterial abundance loadings are given in the right panel of **Figure 9**. The most notable observation is that the first axes places more weight on rarer species, while the second places proportionally more weight on abundant species. Further, the two axes seem to have very different behaviors with respect to Prevotellaceae and Veillonellaceae.

In general, we find the results from the LDA–CCA approach less satisfying than those of the sparse CCA of section Penalized Matrix Decomposition. It seems that inference of a probabilistic model with shared and unshared parameters is more difficult than optimization of a single set of shared parameters. It may be possible to improve this approach through the following strategies:

- Applying LDA–CCA only to those species that are not sent entirely to zero by sparse CCA.
- Placing a sparsity-inducing prior on the scores B^X, B^Y, W^X, and W^Y, respectively, in the spirit of Archambeau and Bach (2009).

Curds & Whey

The Curds & Whey (C&W) procedure is a "soft" version of reduced-rank regression, differentially shrinking the ordinary least squares (OLS) fits with respect to the response canonical correlation directions (Breiman and Friedman, 1997). This is in contrast to reduced-rank regression, whose projection onto the first K response canonical correlation directions is a hard-thresholding analog. Hence, C&W is to reduced-rank regression what ridge regression is to principal component regression.

More precisely, the C&W algorithm fits a table Y according to

$$\hat{Y} = P_X Y V \Lambda^{-1}, \tag{10}$$

where again $V \in \mathbb{R}^{p_1 \times p_1}$ are the CCA directions associated with the response Y and P_x is the projection operator onto the column space of X. Λ is defined to be a diagonal matrix that determines the degree of shrinkage for the different canonical directions.

The main difficulty in C&W is the choice of Λ, and Breiman and Friedman (1997) suggest several possibilities. One choice is derived from a generalized cross-validation point of view, and results in shrinkage towards the response canonical correlation directions, without assuming the form of equation (10) *a priori*. This derivation is provided in section *Derivation of Curds & Whey Shrinkage*.

Graph-Fused Lasso

An approach to multiresponse regression, introduced by Chen et al. (2010), incorporates prior knowledge about the relationship between responses. Specifically, they use the correlation network between responses to induce structured regularization on the regression parameters.

Let $Y \in \mathbb{R}^{n \times p_1}$ and $X \in \mathbb{R}^{n \times p_2}$ and assume a correlation network between the p_2 tasks. This is denoted by $G = (V, E)$, where $V = \{1,\ldots,p_1\}$. Each edge e is associated with a weight, $r(e)$, giving the correlation between the pair of responses.

The graph-fused lasso estimates a coefficient matrix $B \in \mathbb{R}^{p_2 \times p_1}$ whose columns $\beta^{(r)}$ are the regression coefficients across tasks, but which have been pooled together, with the strength of the pooling depending on the separately computed strength of the relationship between tasks. Formally, $\hat{\beta}$ is defined as the solution to the optimization,

$$\underset{B \in \mathbb{R}^{p_2 \times p_1}}{\text{minimize}} \frac{1}{2}||Y - XB||_F^2 + \lambda||B||_1 + \gamma \sum_{e \in E} \sum_{j=1}^{p_2} |r_e| \left| \beta_j^{(e+)} - \text{sign}(r_e)\beta_j^{(e-)} \right|, \tag{11}$$

where $||B||_1$ is the sum of the absolute values of all entries of B, β_j is the j^{th} row of B, and e^- and e^+ denote the nodes at either end of the edge e. The last regularization term in the objective is called the graph fused-lasso penalty, and it is this element that encourages pooling of information across regression problems.

Example

We apply the graph-fused lasso to the body composition problem and compare it to a naive version of the lasso that does not share any information across responses. We consider predicting the body composition variables, many of which are strongly correlated with one another, using variance-stabilized bacterial abundances.

We filter away species that do not appear in at least 7% of samples, as in the original PCA approach. We set the smoothing parameter to $\mu = 0.01$, while the ℓ^1 and graph-regularization parameters are set to $\lambda = 0.1$ and $\gamma = 0.01$, respectively, after they were heuristically found to provide interpretable levels of sparsity and smoothness in the fitted coefficients.

The graph-fused lasso requires a correlation graph between response variables. We estimate such a graph using the graphical lasso (Friedman et al., 2008), since there are only ~100 with which to estimate the 36-dimensional covariance matrix. The estimated correlation matrix is displayed in **Figure 11**.

The fitted coefficients from the graph-fused lasso are given in the top panel of **Figure 12**. The analogous display when the problem is decoupled into parallel lasso regressions is given in the bottom panel of the same figure.

Generally, both approaches highlight the same directions and size of association between individual species and the response variables, though those returned by the graph-fused lasso are smoother across responses. This smoothing may obscure true variation—for example, the stronger association between height_dxa and a few Ruminoccocus species—that appears in the parallel-lasso approach. On the other hand, regularization reduces the number of one-off nonzero coefficients, which are likely just noise.

There appear to be real associations between Lachnospiraceae and Ruminococcaceae and the body composition measurements. The strongest negative association between species abundance and fat mass occurs among a few species of Ruminococcaceae. Most species that have any association tend to have the same direction and magnitude of association across all body composition variables, not just those restricted to one mass type. This seems to be the case even in the parallel-lasso context, where such structure has not been directly imposed.

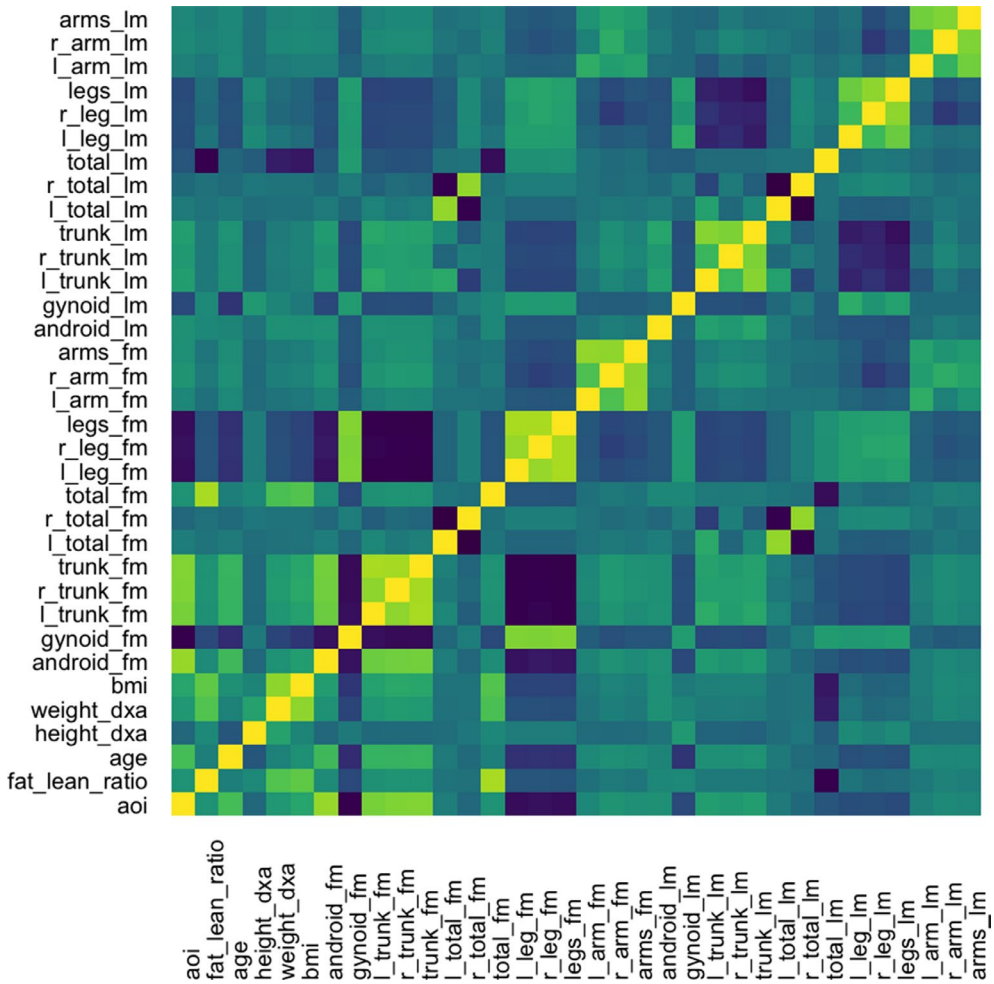


FIGURE 11 | Correlation matrix used as the input graph R for the graph-fused lasso, estimated itself according to the graphical lasso.

DISCUSSION

In this work, we have studied the problem of multitable data analysis, reviewing both the algorithmic foundations and practical applications of various methods. We have described approaches that are usually confined to particular literature areas and highlighted certain similarities in the process—for example, PCA-IV (section PCA-IV) and the graph-fused lasso (section Graph-Fused Lasso) were proposed in very different contexts, but have similar goals. By writing short, self-contained descriptions of various methods, we hope to contribute to an effort to distill ideas from the wide multitable data analysis literature to make them easily understandable to researchers interested in entering this field and useful for scientists hoping to apply these methods. A "cheat-sheet" summarizing some of the key properties of these methods is given in **Table 1**, and relevant packages can be found in **Table 2.**

In developing our WELL-China case study, we have both 1) described the types of interpretations facilitated by different approaches and 2) provided accessible implementations that can be incorporated into practical scientific workflows. Though our focus on a single application has allowed side-by-side comparisons of methods, we do not want to leave the reader with the impression that these methods are tied in any way to this particular biological analysis task. Indeed, the value of mathematical abstractions is that they can be applied to situations outside the imaginations of the original method designers. For example, consider these potential use cases:

- *Microbiome and metabolites:* If we replace the body composition table with the concentrations of different metabolites across samples, we can begin to make claims about covariation between microbiome community composition and host metabolic processes (Chong and Xia, 2017; Fukuyama et al., 2017).
- *Microbiome and metagenomics:* In addition to a species composition matrix, we might have data quantifying the presence of various genes. The methods in this review could be used to understand the relationship between community composition and functional capacity (Gill et al., 2006; Kurokawa et al., 2007).
- *Microbiome and perturbations:* If we had a matrix tracking the application of various perturbations to the host—the use of various medications, for example—we could use

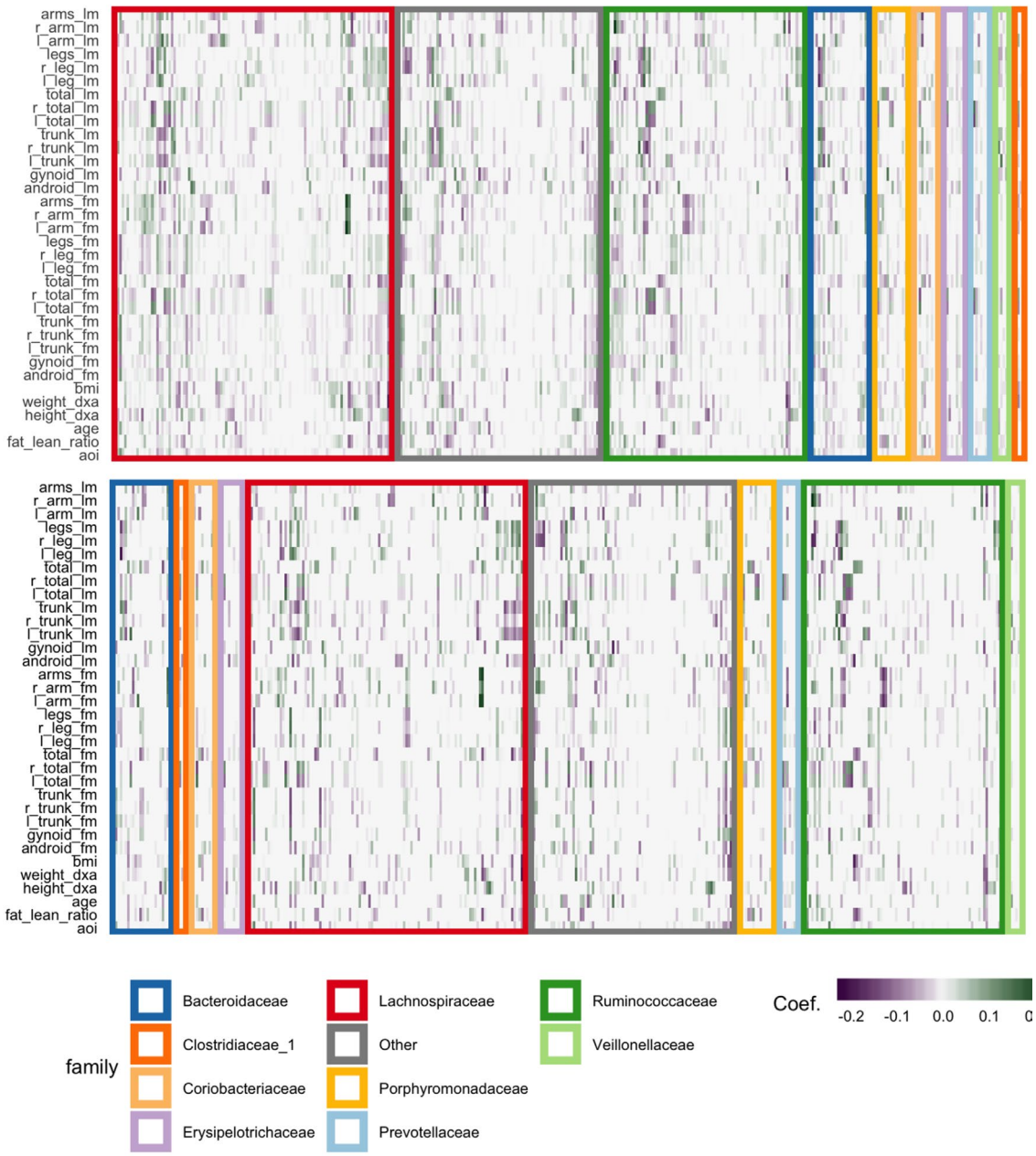


FIGURE 12 | Coefficients for the graph-fused (top) and decoupled (bottom) lasso fits highlight groups of species with similar profiles across response variables. Colored rectangles demarcate taxonomic families. Individual cells give the coefficient for a particular species (column) for a given response variable (row). Purple and green denote negative and positive coefficients, respectively. Note that coefficient graph-fused panels have been smoothed according to correlation network between variables, as given in **Figure 11**. Species with similar coefficients are placed near one another. Note that even in the decoupled case, where there is no sharing across response problems, the coefficients nonetheless seem to be similar within lean and fat mass response groups, respectively. However, they are not as smooth as in the graph-fused lasso. As there is some consistency within these groups of variables, the form of structured regularization imposed by the graph seems appropriate.

multitable methods to describe ways these (multidimensional) perturbations are related to microbiome community structure (Dethlefsen and Relman, 2011).

Our case study includes carefully thought-through visualizations of model results, a step that is crucial in scientific study but often overlooked in methodological research, where model results are reduced to tables of performance metrics. Recognizing that a good deal of effort in statistical work goes into data preparation and visualization of model results, we have ensured that codes for all steps are available, so that our work is fully reproducible.

TABLE 1 | A high-level comparison of the multitable analysis methods discussed in this review. The purpose of this table is to give rules-of-thumb that can guide practical application, where choices invariably depend on the scale and structure of the data, the goals of the analysis, the expected number of future workflow applications, and availability of programming computation time.

Property	Algorithms	Consequence
Analytical solution	Concat. PCA, CCA, ColA, MFA, PTA, Statico/Costatis	Methods with analytical solutions generally run much faster than those that require iterative updates, optimization, or Monte Carlo sampling. They tend to be restricted to more classical settings, however.
Require covariance estimate	Concat. PCA, CCA, ColA, MFA, PTA, Statico/Costatis	Methods that require estimates of covariance matrices cannot be applied to data with more variables than samples, and become unstable in high-dimensional settings.
Sparsity	SPLS, Graph-Fused Lasso, Graph-Fused Lasso	Encouraging sparsity on scores or loadings can result in more interpretable, results for high-dimensional data sets. These methods provide automatic variable selection in the multitable analysis problem.
Tuning parameters	*Sparsity*: Graph-Fused Lasso, PMD, SPLS *Number of Factors*: PCA-IV, Red. Rank Regression, Mixed-Membership CCA Prior *Parameters*: Mixed- Membership CCA, Bayesian Multitask Regression	Methods with many tuning parameters are often more expressive than those without any, since it makes it possible to adapt to different degrees of model complexity. However, in the absence of automatic tuning strategies, these methods are typically more difficult to use effectively.
Probabilistic	Mixed-Membership CCA, Bayesian Multitask Regression	Probabilistic techniques provide estimates of uncertainty, along with representations of cross-table covariation. This comes at the cost of more involved computation and difficulty in assessing convergence.
Not Normal or Nonlinear	CCpNA, Mixed-Membership CCA, Bayesian Multitask Regression	When data are not normal (and are difficult to transform to normality) or there are sources of nonlinear covariation across tables, it can be beneficial to directly model this structure.
>2 Tables	Concat. PCA, CCA, MFA, PMD	Methods that allow more than two tables are applicable in a wider range of multitable problems. Note that these are a subset of the cross-table symmetric methods.
Cross-Table Symmetry	Concat. PCA, CCA, ColA, Statico/Costatis, MFA, PMD	Cross-table symmetry refers to the idea that some methods don't need a supervised or multitask setup, where one table contains response variable and the other requires predictors. The results of these methods do not change when the two tables are swapped in the method input.

TABLE 2 | Pointers to R package that can be used to implement methods discussed in this survey. The vignettes in these packages go into more depth on the capabilities of these packages than do the short scripts used in our case study, available at https://github.com/krisrs1128/multitable_review.

Package	Methods	Documentation	Link
ade4	PCA, CCA, ColA, Statico, Costatis, PCA-IV	Average	https://cran.r-project.org/web/packages/ade4/
FactoMineR	PCA, MFA	High	https://cran.r-project.org/web/packages/FactoMineR/
vegan	CCA, CCpnA	High	https://cran.r-project.org/web/packages/vegan/
spls	SPLS	High	https://cran.r-project.org/web/packages/spls/
PMA	PMD	High	https://cran.r-project.org/web/packages/PMA/
pls	PLS	High	https://cran.r-project.org/web/packages/pls/
Base R	PCA, CCA	High	https://cran.r-project.org/
GFLasso	Graph-Fused Lasso	Low	https://github.com/krisrs1128/gflasso
bayesMult	Bayesian Multitask Regression	Low	https://github.com/krisrs1128/bayesmult

We have found that multitable data analysis problems have motivated a wide range of analysis approaches. This is not surprising, considering the variety of contexts in which it arises, and it speaks to the richness of this methodological problem. As new data sources arise and as science evolves, we expect these ideas will inspire future generations of multitable research advances.

AUTHOR CONTRIBUTIONS

SH and KS conceived and designed the review, drafted the manuscript, and prepared all figures. KS implemented code for data analysis.

FUNDING

KS was supported by a Stanford University Weiland fellowship and the National Institutes of Health T32 grant 5T32GM096982-04. SH is supported by the National Institutes of Health TR01 grant AI112401.

ACKNOWLEDGMENTS

We thank the WELL-China study team for sharing the data appearing in this study and Yan Min for useful discussions.

An earlier version of this work first appeared in KS's PhD thesis (Sankaran, 2018).

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi: 10.1186/gb-2010-11-10-r106

Archambeau, C., and Bach, F. R., (2009). Sparse probabilistic projections. In *Advances in neural information processing systems*. 73–80.

Ashish, N., Ambite, J. L., Muslea, M., and Turner, J. A. (2010). Neuroscience data integration through mediation: an (f) birn case study. *Front. Neuroinform.* 4, 118. doi: 10.3389/fninf.2010.00118

Bach, F. R., and Jordan, M. I. (2005). *A probabilistic interpretation of canonical correlation analysis*. Berkeley: Technical Report 688 Department of Statistics, University of California.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993

Breiman, L., and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. R. Stat. Soc. Series B Stat. Methodol.* 59, 3–54. doi: 10.1111/1467-9868.00054

Buhlmann, P., and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Berlin Heidelberg: Springer Science & Business Media. doi: 10.1007/978-3-642-20192-9

Chalise, P., and Fridley, B. L. (2017). Integrative clustering of multi-level 'omic' data based on non-negative matrix factorization algorithm. *PLOS ONE* 12, e0176278. doi: 10.1371/journal.pone.0176278

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X., (2017). Deep learning based multi-omics integration robustly predicts survival in liver cancer. *Clin*. Cancer Res. 24 (6), 1248–1259. doi: 10.1101/114892

Chen, X., Kim, S., Lin, Q., Carbonell, J. G., and Xing, E. P. (2010). Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*. https://arxiv.org/abs/1005.3579

Chong, J., and Xia, J. (2017). Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites* 7, 62. doi: 10.3390/metabo7040062

Chun, H., and Kele, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* 72, 3–25. doi: 10.1111/j.1467-9868.2009.00723.x

Chung, D., Chun, H., and Keles, S. (2012). Spls: Sparse partial least squares (spls) regression and classification. *R package, version* 2, 1–1.

Dethlefsen, L., and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci.* 108, 4554–4561. doi: 10.1073/pnas.1000087107

Dolédec, S., and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshw. Biol.* 31, 277–294. doi: 10.1111/j.1365-2427.1994.tb01741.x

Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–135. doi: 10.1080/00401706.1993.10485033

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451

Friedman, J., Hastie, T., and Tibshirani, R., (2001). *The elements of statistical learning* Vol. 1. Berlin: Springer series in statistics Springer. doi: 10.1007/978-0-387-21606-5_1

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Fukuyama, J., Rumker, L., Sankaran, K., Jeganathan, P., Dethlefsen, L., Relman, D. A., et al. (2017). Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.* 13, e1005706. doi: 10.1371/journal.pcbi.1005706

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* 8, I1. doi: 10.1186/1752-0509-8-S2-I1

Greenacre, M. J. (1984). Theory and applications of correspondence analysis. *J. Am. Stat. Assoc.* 82 (398), 437–447.

Greenacre, M., and Hastie, T. (1987). The geometric interpretation of correspondence analysis. *J. Am. Stat. Assoc.* 82, 437–447. doi: 10.1080/01621459.1987.10478446

Gustafsson, M. G. (2001). A probabilistic derivation of the partial least-squares algorithm. *J. Chem. Inf. Comput. Sci.* 41, 288–294. doi: 10.1021/ci0003909

Hannan, E. (1967). Canonical correlation and multiple equation systems in economics. *Econometrica*, 35(1), 123–138. doi: 10.2307/1909387

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.1093/biomet/28.3-4.321

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *J. Comput. Graph. Stat.* 12, 531–547. doi: 10.1198/1061860032148

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken: John Wiley & Sons. doi: 10.1002/9780470238004

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., et al. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 14, 169–181. doi: 10.1093/dnares/dsm018

Lê Cao, K.-A., Rossouw, D., Robert-Granie, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* 7, 1544–6115. doi: 10.2202/1544-6115.1390

Lee, D. D., and Seung, H. S. (2001). "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*. Eds. T. K. Leen, T. G. Dietterich, and V. Tresp (Cambridge, MA: MIT Press), 556–562.

Ley, R. E. (2010). Obesity and the human microbiome. *Curr. Opin. Gastroenterol.* 26, 5–11. doi: 10.1097/MOG.0b013e328333d751

Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11070–11075. doi: 10.1073/pnas.0504978102

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Mardia, K. V., Kent, J. T., and Bibby, J. M., (1980). Multivariate analysis. London: Academic Press

Matsuzawa, Y. (2008). The role of fat topology in the risk of disease. *Int. J. Obes.* 32, S83. doi: 10.1038/ijo.2008.243

McHardy, I. H., Goudarzi, M., Tong, M., Ruegger, P. M., Schwager, E., Weger, J. R., et al. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 1, 17. doi: 10.1186/2049-2618-1-17

Min, Y., Ma, X., Sankaran, K., Ru, Y., Chen, L., Baiocchi, M., et al. (2019). Sex-specific association between gut microbiome and fat distribution. *Nat. Commun.* 10, 2408. doi: 10.1038/s41467-019-10440-5

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009). Bayesian exponential family pca proceedings of advances in neural information processing systems. *Adv. Neural. Inf. Process. Syst.* 1089–1096.

Pagés, J. (2014). *Multiple Factor Analysis by example using R*. CRC Press. doi: 10.1201/b17700

Pagés, J., and Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemom. Intell. Lab. Syst.* 58, 261–273. doi: 10.1016/S0169-7439(01)00165-4

Pagés, J. (2004). Multiple factor analysis: main features and application to sensory data. *Rev. Colomb. Estad.* 27 (1), 1.

Perez, P., and de Los Campos, G. (2014). Genome-wide regression & prediction with the bglr statistical package. *Genetics*. 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Rahnavard, G., Franzosa, E. A., Mclver, L. J., Schwager, E., Weingart, G., Moon, Y. S. et al., (2017). High-sensitivity pattern discovery in large multi'omic datasets. http://huttenhower.sph.harvard.edu/halla

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhā*, 26 (4), 329–358. https://www.jstor.org/stable/25049339

Sankaran, K. (2018). Discovery and visualization of latent structure with applications to the microbiome. Ph.D. thesis, Stanford University.

Stone, M., and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Stat. Soc. Series B Stat. Methodol.*52 (2), 237–269. doi: 10.1111/j.2517-6161.1990.tb01786.x

Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167–1179. doi: 10.2307/1938672

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *Ann. Appl. Stat.* 5 (4), 2300–2325. doi: 10.1214/10-AOAS372

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480. doi: 10.1038/nature07540

Vlassis, N., Motomura, Y., and Krose, B., (2000). "Supervised linear feature extraction for mobile robot localization," in *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on (IEEE)*, vol. 3, 2979–2984.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 10 (3), 515–534. doi: 10.1093/biostatistics/kxp008

Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). Package 'pma'. *Genet. Mol. Biol.* 8, 28.

Wold, H. (1985). "Partial least squares," in *Encyclopedia of statistical sciences*. Vol. 6. New York: John Wiley, 581–591.

Zhu, M., Hastie, T. J., and Walther, G. (2005). Constrained ordination analysis with flexible response functions. *Ecol. Modell.* 187, 524–536. doi: 10.1016/j.ecolmodel.2005.01.049

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286. doi: 10.1198/106186006X113430

Conflict of Interest Statement: 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.

APPENDIX

This appendix includes derivations and technical discussion of several methods surveyed in the main text: PCA-IV, PTA, and the C&W algorithm. While these methods can be understood and applied based on their computational description, these mathematical discussions provide motivation and context for their particular form.

DERIVATION DETAILS FOR PCA-IV

In this section, we provide the argument for why the generalized eigendecomposition $\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} = \hat{\Sigma}_{XX} V \Lambda V^T$ provides the optimal V used in PCA-IV.

First consider $k = 1$. For any $\tilde{v}$, the objective in equation (5) has the form

$$\begin{aligned}\mathrm{tr}\left(\hat{\Sigma}_{YX}\tilde{v}\left(\tilde{v}\hat{\Sigma}_{XX}\tilde{v}\right)^{-1}\left(\hat{\Sigma}_{YX}\tilde{v}\right)^{T}\right) &= \frac{\tilde{v}^T\hat{\Sigma}_{XY}\hat{\Sigma}_{YX}\tilde{v}}{\tilde{v}^T\hat{\Sigma}_{XX}\tilde{v}} \\ &= \frac{\tilde{w}^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}^{-\frac{1}{2}}\Sigma_{YX}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\tilde{w}}{||\tilde{w}||_2^2}\end{aligned} \tag{12}$$

where we change variables $\tilde{w} = \Sigma_{XX}^{\frac{1}{2}}\tilde{v}$. But to maximize equation (12), just choose $\tilde{w}$ to be the top eigenvector of $\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}$, which implies that $\tilde{v}$ is the top generalized eigenvector of $\Sigma_{XY}\Sigma_{YX}$ with respect to Σ_{XX}. Indeed, in this case,

$$\begin{aligned}\Sigma_{XY}\Sigma_{YX}\tilde{v} &= \Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\tilde{w} \\ &= \Sigma_{XX}^{\frac{1}{2}}\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\tilde{w} \\ &= \Sigma_{XX}^{\frac{1}{2}}\lambda_1\tilde{w} \\ &= \lambda_1\Sigma_{XX}\tilde{v}.\end{aligned}$$

Hence, in the case $K = 1$, the criterion is maximized by the top generalized eigenvector. For larger K, recall that the problem of maximizing $\frac{v^TAv}{||v||^2}$ over v subject to being orthogonal to the first $K-1$ eigenvectors of A is solved by the K^{th} eigenvector of A, and applying this fact in step 12 of the argument above gives the result for general K.

DERIVATION OF PTA α

The Lagrangian of the optimization defined by PTA is

$$\mathcal{L}(\alpha,\lambda) = \sum_{l=1}^{L}\alpha_l\left\langle\bar{X}, X_{..l}\right\rangle + \lambda(||\alpha||_2^2 - 1),$$

Which, when differentiated with respect to α, yields $\alpha_l = -\frac{1}{2\lambda}\left\langle\bar{X}, X_{..l}\right\rangle$ for all l. The constraint that $||\alpha||_2^2 = 1$ implies that $\frac{1}{4\lambda^2}\sum_{l'=1}^{L}\left\langle\bar{X}, X_{..l'}\right\rangle^2 = 1$, which gives $\lambda = \frac{1}{2}\sqrt{\sum_{l'=1}^{L}\left\langle\bar{X}, X_{..l'}\right\rangle^2}$, so $\alpha_l = \frac{\left\langle\bar{X}, X_{..l}\right\rangle}{\sqrt{\sum_{l'=1}^{L}\left\langle\bar{X}, X_{..l'}\right\rangle^2}}$.

DERIVATION OF CURDS & WHEY SHRINKAGE

Consider prediction across many related response variables. One way to pool information across responses is to define new fitted values from a linear combination of independent OLS fits. That is, to predict a response $y_i \in \mathbb{R}^{p_1}$, we set $\hat{y}_i^{cw} = B\hat{y}_i^{\text{ols}}$ for some square matrix $B \in \mathbb{R}^{p_1 \times p_1}$. But how to choose B?

One reasonable idea is to choose a B that has the best performance in a generalized cross-validation (GCV). The GCV approximation is that the h_{ii} can be approximated by their average across all diagonal elements of H: $h_{ii} \approx h := \frac{1}{n}\mathrm{tr}(H)$ for all i. In this spirit, define $g = \frac{1}{1-h}$ and approximate

$$\hat{y}_{-1} \approx (1-g)y_i + g\hat{y}_i$$

Then, the leave-one-out CV error can be simplified to

$$\sum_{i=1}^{n}||y_1 - B\hat{y}_{-i}||_2^2 = \sum_{i=1}^{n}||y_i - B((1-g)y_i + g\hat{y}_{-i}||_2^2,$$

and differentiating with respect to B, we find that the optimal $\hat{B}^{\text{cw}}$ in this GCV framework must satisfy

$$\sum_{i=1}^{n}(y_i - B((1-g)y_i + g\hat{y}_{-i}))((1-g)y_i + g\hat{y}_{-i})^T,$$

or equivalently

$$\sum_{i=1}^{n}y_i((1-g)y_i + g\hat{y}_{-i})^T = \sum_{i=1}^{n}B((1-g)y_i + g\hat{y}_{-1})((1-g)y_i + g\hat{y}_{-1})^T,$$

which in matrix form is

$$(1-g)Y^TY + g\hat{Y}^TY = B\left((1-g)Y_i + g\hat{Y})\right)^T\left((1-g)Y_i + g\hat{Y})\right), \tag{13}$$

where $\hat{Y} \in \mathbb{R}^{n \times p_1}$ has i^{th} row $\hat{y}_{-i}$.

Next, we can represent these cross-products in a way that is suggestive of CCA,

$$
\begin{aligned}
Y^T Y &= n\hat{\Sigma}_{YY} \\
\hat{Y}^T Y &= Y^T H Y = Y^T X\left(X^T X\right)^{-1} X^T Y = n\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY} \\
\hat{Y}^T \hat{Y} &= Y^T P_X^2 Y = Y^T P_X Y = n\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY},
\end{aligned}
$$

Substituting this into equation (13) and ignoring the scaling n yields

$$
\begin{aligned}
&(1-g)\hat{\Sigma}_{YY} + g\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY} = \\
&B\left[(-g)\hat{\Sigma}_{YY} + (2g-g^2)\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY}\right].
\end{aligned}
$$

Postmultiplying by $\hat{\Sigma}_{YY}$ gives

$$
(1-g)I_{p1} + g\hat{Q}^T = B[(1-g)I_{p1} + (2g-g^2)\hat{Q}^T], \qquad (14)
$$

where

$$
\hat{Q} := \hat{\Sigma}_{YY}^{-1}\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY} \in \mathbb{R}^{p_1 \times p_1}
$$

Now, we claim that we can decompose $\hat{Q} = VD^2V^{-1}$, where $V \in \mathbb{R}^{p_1 \times p_1}$ is the full matrix of CCA response directions and D is diagonal with the canonical correlations. Indeed, the usual CCA response directions V can be recovered by setting $V = \hat{\Sigma}_{YY}^{-\frac{1}{2}}\tilde{V}$, where $\tilde{V}$ comes from the SVD of $A := \Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{XX}^{-\frac{1}{2}} = \tilde{U}D\tilde{V}^T$. Hence

$$
\begin{aligned}
Q &= \Sigma_{YY}^{-\frac{1}{2}} A^T A \Sigma_{YY}^{\frac{1}{2}} \\
&= \Sigma_{YY}^{-\frac{1}{2}} \tilde{V}^2 D^2 \tilde{V}^T \Sigma_{YY}^{\frac{1}{2}} \\
&= VD^2V^{-1},
\end{aligned}
$$

where we are able to write $V^{-1} = \tilde{V}^T\Sigma_{YY}^{-\frac{1}{2}}$ because $\tilde{V}$ is the full (untruncated) matrix of eigenvectors, so $\tilde{V}\tilde{V}^T = I$ in addition to the usual $\tilde{V}^T\tilde{V} = I$, which holds even for the truncated SVD.

Therefore, equation (14) can be expressed as

$$
V^{-T}[(1-g)]I_{p_1} + gD^2]V^T = BV^{-T}[(1-g)]I_{p_1} + (2g-g^2)\mathrm{D}^2]V^T
$$

and the B satisfying the normal equations has the form

$$
\hat{B}^{\mathrm{cw}} = V^{-T}\Lambda VT,
$$

where Λ is a diagonal matrix with entries

$$
\lambda_{jj} = \frac{1-g+d_{jjg}^2}{1-g+(2g-g^2)d_{jj}^2}
$$

Notice that when n is large, $\frac{1}{n}\mathrm{tr}\,P_X$ will be small, leading to a smaller $g \approx 0$ and less shrinkage. Recall that $\hat{B}^{\mathrm{cw}}$ is used to pool across OLS fits, $\hat{y}_i^{\mathrm{cw}} = \hat{B}^{\mathrm{cw}}\hat{y}_i^{ols}$. That is,

$$
\hat{Y}^{\mathrm{cw}} = \hat{Y}^{\mathrm{ols}}B^T = \hat{Y}^{\mathrm{ols}}V\Lambda V^{-1}
$$

which we can also view as $\hat{Y}^{\mathrm{cw}}V = \left(\hat{Y}^{\mathrm{ols}}V\right)\Lambda$. This means that the C&W coordinates along the canonical directions V are set as the OLS fits $\hat{Y}^{\mathrm{ols}}$ along the canonical directions V, with weights defined by Λ. The actual $\hat{Y}^{\mathrm{cw}}$ are recovered by transforming back to the original coordinate system. A similar way to view the C&W fits is to note $\hat{Y}^{\mathrm{cw}}V = P_X(YV)\Lambda$, which is the original data Y according to the canonical directions, then projects the shrunk data onto the subspace defined by the columns of X. In any case, we see that C&W pools across regression problems through a soft shrinkage weighted along canonical response directions.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 07 November 2019
doi: 10.3389/fgene.2019.00963

A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types

Antoine Bodein[1†], Olivier Chapleur[2†], Arnaud Droit[1] and Kim-Anh Lê Cao[3]*

[1] Molecular Medicine Department, CHU de Québec Research Center, Université Laval, Québec, QC, Canada, [2] Hydrosystems and Biopresses Research Unit, Irstea, Antony, France, [3] Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC, Australia

Simultaneous profiling of biospecimens using different technological platforms enables the study of many data types, encompassing microbial communities, omics, and meta-omics as well as clinical or chemistry variables. Reduction in costs now enables longitudinal or time course studies on the same biological material or system. The overall aim of such studies is to investigate relationships between these longitudinal measures in a holistic manner to further decipher the link between molecular mechanisms and microbial community structures, or host-microbiota interactions. However, analytical frameworks enabling an integrated analysis between microbial communities and other types of biological, clinical, or phenotypic data are still in their infancy. The challenges include few time points that may be unevenly spaced and unmatched between different data types, a small number of unique individual biospecimens, and high individual variability. Those challenges are further exacerbated by the inherent characteristics of microbial communities-derived data (e.g., sparse, compositional). We propose a generic data-driven framework to integrate different types of longitudinal data measured on the same biological specimens with microbial community data and select key temporal features with strong associations within the same sample group. The framework ranges from filtering and modeling to integration using smoothing splines and multivariate dimension reduction methods to address some of the analytical challenges of microbiome-derived data. We illustrate our framework on different types of multi-omics case studies in bioreactor experiments as well as human studies.

Keywords: time course, data integration, splines, feature selection, dimension reduction, multi-omics

OPEN ACCESS

Edited by:
Himel Mallick, Merck, United States

Reviewed by:
Gholamali Ali Rahnavard, Broad Institute, United States
Lingling An, University of Arizona, United States

***Correspondence:**
Kim-Anh Lê Cao
kimanh.lecao@unimelb.edu.au

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 April 2019*
Accepted: *10 September 2019*
Published: *07 November 2019*

Citation:
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

INTRODUCTION

Microbial communities are highly dynamic biological systems that cannot be fully investigated in snapshot studies. The decreasing cost of DNA sequencing has enabled longitudinal and time-course studies to record the temporal variation of microbial communities (Knight et al., 2012; Faust et al., 2015). These studies can inform us about the stability and dynamics of microbial communities in response to perturbations or different conditions of the host or their habitat. They can also capture the dynamics of microbial interactions (Bucci et al., 2016; Ridenhour et al., 2017) or associated

changes of microbial features, such as taxonomies or genes, to a phenotypic group (Metwally et al., 2018).

However, besides the inherent characteristics of microbiome data, including sparsity, compositionality (Aitchison, 1982; Gloor et al., 2017), its multivariate nature, and high variability (Lê Cao et al., 2016a), longitudinal studies suffer from irregular sampling and subject drop-outs. Thus, appropriate modeling of the microbial profiles is required—for example, by using spline modeling. Methods including loess (Shields-Cutler et al., 2018), smoothing spline ANOVA (Paulson et al., 2017), negative binomial smoothing splines (Metwally et al., 2018), or Gaussian cubic splines (Luo et al., 2017) were proposed to model dynamics of microbial profiles across groups of samples or subjects. The aim of these approaches is to make statistical inferences about global changes of differential abundance across multiple phenotypes of interest, rather than at specific time points. These proposed methods are univariate and, as such, cannot infer ecological interactions (Morris et al., 2016). Other types of methods aim to cluster microbial profiles to posit hypotheses about symbiotic relationships, interaction, or competition. For example, Baksi et al. (2018) used a Jenson–Shannon divergence metric to visually compare metagenomic time series.

Multivariate ordination methods can exploit the interaction between microorganisms but need to be used with sparsity constraints, such as ℓ_1 regularization (Tibshirani, 1996), to reduce the number of variables and improve interpretability through variable selection. Several sparse methods were proposed and applied to microbiome studies, such as sparse linear discriminant analysis (Clemmensen et al., 2011) and sparse partial least squares discriminant analysis (sPLS-DA, Lê Cao et al., 2016b), but for a single time point. Therefore, further developments are needed to combine time-course modeling with multivariate approaches to start exploring microbial interactions and dynamics.

In addition, current statistical methods have mainly focused on a single microbiome dataset, rather than the combination of different layers of molecular information obtained with parallel multi-omics assays performed on the same biological samples. Data derived from each omics technique are typically studied in isolation and disregard the correlation structure that may be present between the multiple data types. Hence, integrating these datasets enables us to adopt a holistic approach to elucidate patterns of taxonomic and functional changes in microbial communities across time. Some sparse multivariate methods have been proposed to integrate omics and microbiome datasets at a single time point and identify sets of features (multi-omics signatures) across multiple data types that are correlated with one another. For example, Gavin et al. (2018) used the DIABLO method (Singh et al., 2019) to integrate 16S amplicon microbiome, proteomics, and metaproteomics data in a type I diabetes study; Guidi et al. (2016) used sparse PLS (Lê Cao et al., 2008) to integrate environmental and metagenomic data from the Tara Oceans expedition to understand carbon export in oligotrophic oceans, and Fukuyama et al. (2017) used sparse canonical correlation analysis (Witten et al., 2009) to integrate 16S and metagenomic data. However, methods or frameworks to integrate multiple longitudinal datasets including microbiome data remain incomplete. Zhou et al. (2008) used principal component analysis (PCA) to summarize functional data, with the PC scores used for model fitting, prediction, and inference. However, only pairwise relationships were investigated and for a single type of data. Other type of modeling (loess regression) was used by Ribicic et al. (2018) in combination with sparse PCA to explore the link between chemistry and microbial community data in the biodegradation of chemically dispersed oil, but their approach was not designed to seek for multi-omics signatures.

We propose a computational approach to integrate microbiome data with multi-omics datasets in longitudinal studies. Our framework, described in **Figure 1** includes smoothing splines in a linear mixed model framework to model profiles across groups of samples and builds on the ability of sparse multivariate ordination methods to identify sets of variables highly associated across the data types, and across time. Our framework encompasses data pre-processing, modeling, data clustering, and integration. It is highly flexible in handling one or several longitudinal studies with a small number of time points, to identify groups of taxa with similar behavior over time and posit novel hypotheses about symbiotic relationships, interactions, or competitions in a given condition or environment, as we illustrate in two case studies.

METHOD

Our proposed approach includes pre-processing for microbiome data, spline modelization within a linear mixed model framework, and a multivariate analysis for clustering and data integration (**Figure 2**).

Pre-Processing of Microbiome Data

We assume the data are in raw count formats resulting from bioinformatics pipelines such as QIIME (Caporaso et al., 2010) or FROGS (Escudié et al., 2017) for 16S amplicon data. Here, we consider the operational taxonomic unit (OTU) level, but other levels can be considered, as well as other types of microbiome-derived data, such as whole genome shotgun sequencing. The data processing step is described in Lê Cao et al. (2016b) and consists of:

1) Low count removal: Only OTUs whose proportional counts exceeded 0.01% in at least one sample were considered for analysis. This step aims to counteract sequencing errors (Kunin et al., 2010).
2) Total sum scaling (TSS) can be considered as a "normalization" process to account for uneven sequencing depth across samples. TSS divides each OTU count by the total number of counts in each individual sample but generates compositional data expressed as proportions. Instead, one can use Centered Log Ratio transformation (CLR), that is scale invariant and addresses in a practical way the compositionality issue arising from microbiome data by projecting the data into a Euclidean space (Aitchison, 1982; Fernandes et al., 2014; Gloor et al., 2017). Given a vector x of p OTU counts for a given sample, CLR

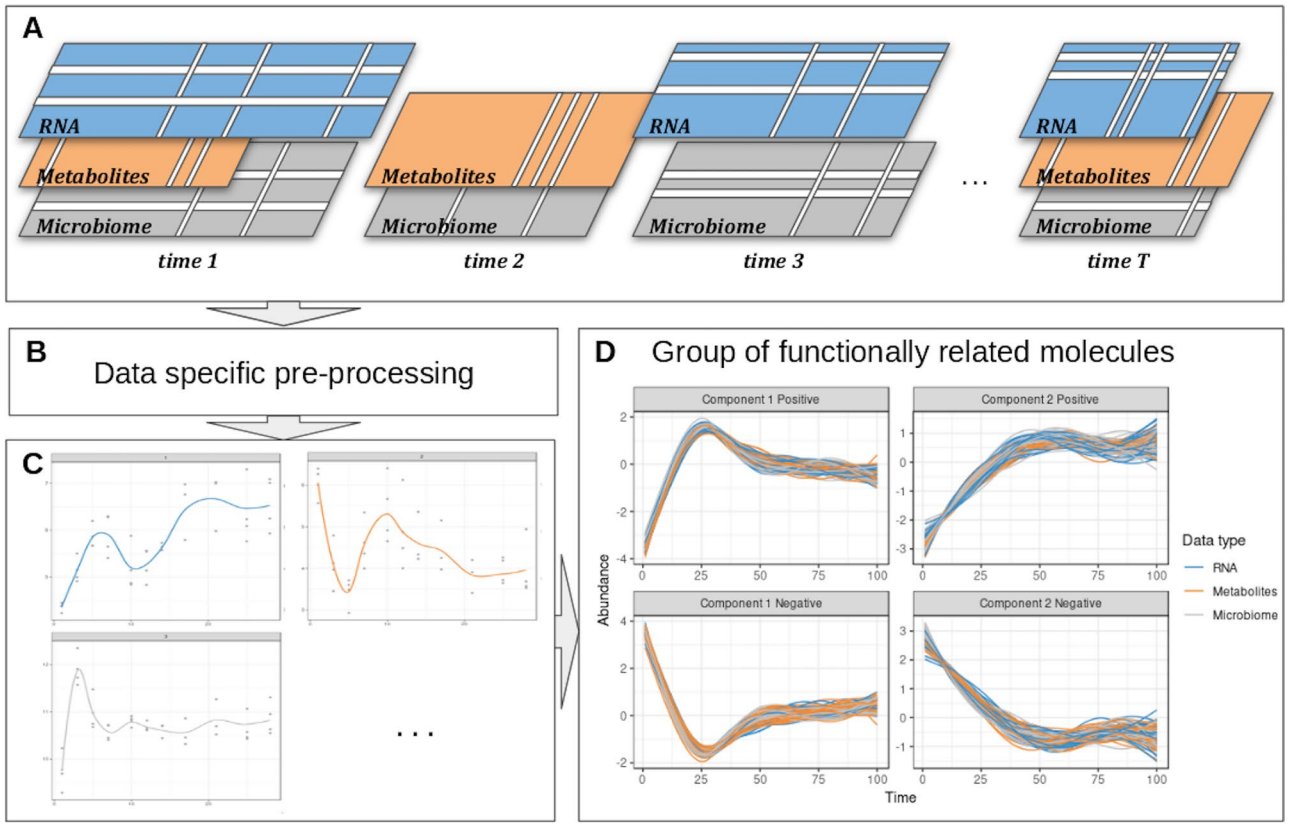


FIGURE 1 | Overview of the proposed approach. **(A)** Description of the experimental design: the same biological material is sampled at several time points across several omic layers indicated in different colors. Blank lines indicate potential missing values per time point or per feature in a given time point. **(B)** Specific pre-processing and normalization are applied according to the type of data. **(C)** Each molecule is modeled as a function of time by taking into account all the variabilities of the different biological replicates in a linear mixed model spline framework. **(D)** The modeled trajectories across all omics layers are clustered using a multivariate integrative method.

(eq. 1) is a log transformation of each element of the vector divided by its geometric mean $G(x)$:

$$\mathrm{clr}(x) = \left[\log\left(\frac{x_i}{G(x)}\right), \ldots, \log\left(\frac{x_p}{G(x)}\right) \right] \quad (1)$$

where

$$G(x) = \sqrt[p]{x_1 \times x_2 \times \ldots x_p}$$

Time Profile Modeling

Linear Mixed Model Splines

The linear mixed model spline (LMMS) modeling approach proposed by Straube et al. (2015) takes into account between and within individual variability and irregular time sampling. LMMS is based on a linear mixed model representation of penalized splines (Durbán et al., 2005) for different types of models. Through this flexible approach of serial fitting, LMMS avoids under- or over-smoothing. Briefly, four types of models are consecutively fitted in our framework on the CLR data:

(1) A simple linear regression of taxa abundance on time, estimated *via* ordinary linear least squares—a straight line that assumes the response is not affected by individual variation
(2) A penalized spline proposed by Durbán et al. (2005) to model nonlinear response patterns
(3) A model that accounts for individual variation with the addition of a subject-specific random effect to the mean response in model (2)
(4) An extension to model (3) that assumes individual deviations are straight lines, where individual-specific random intercepts and slopes are fitted

All four models are described in **Appendix 1.** Straube et al., 2015 showed that the proportion of profiles fitted with the different models increased in complexity with the organism considered. Different types of splines can be considered in models (2)–(4), including a cubic spline basis (Verbyla et al., 1999), a penalized spline and a cubic penalized spline. A cubic spline basis uses all inner time points of the measured time

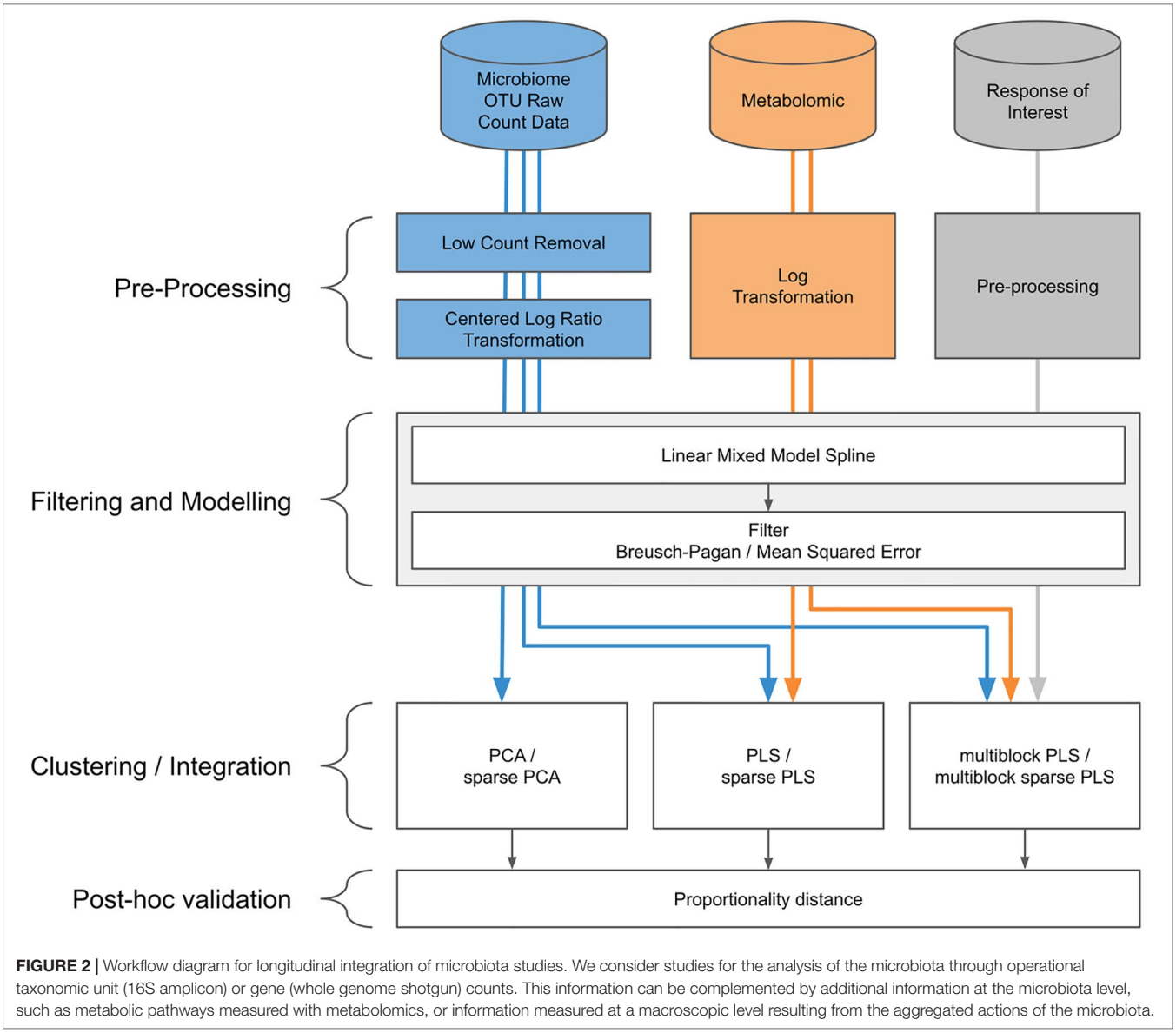


FIGURE 2 | Workflow diagram for longitudinal integration of microbiota studies. We consider studies for the analysis of the microbiota through operational taxonomic unit (16S amplicon) or gene (whole genome shotgun) counts. This information can be complemented by additional information at the microbiota level, such as metabolic pathways measured with metabolomics, or information measured at a macroscopic level resulting from the aggregated actions of the microbiota.

interval as knots and is appropriate when the number of time points is small (≤5), whereas the penalized spline and cubic penalized spline bases use the quantiles of the measured time interval as knots; see Ruppert (2002). In our case studies, we used penalized splines. The LMMS models are implemented in the R package `lmms` (Straube et al., 2016).

Prediction and Interpolation

The fitted splines enable us to predict or interpolate time points that might be missing within the time interval (e.g., inconsistent time points between different types of data or covariates). Additionally, interpolation is useful in our multivariate analyses described below to smooth profiles, and when the number of time points is small (≤5). In the following section, we therefore consider data matrices $\boldsymbol{X}$ $(T \times P)$, where T is the number of (interpolated) time points and P the number of taxa. The individual dimension has thus been summarized through the spline fitting procedure, so that our original data matrix of size $(N \times P \times T)$, where N is the number of biological samples, is now of size $(T \times P)$.

Filtering Profiles After Modeling

A simple linear regression model (1) might be the result of highly noisy data. To retain only the most meaningful profiles, the quality of these models was assessed with a Breusch–Pagan test to indicate whether the homoscedasticity assumption of each linear model was met (Breusch and Pagan, 1979) simple. We also used a threshold based on the mean squared error (MSE) of the linear models, by only including profiles for which their MSE was below the maximum MSE of the more complex fitted models (2)–(4). The latter filter was only applied when a large number of linear models (1) were fitted and the Breusch–Pagan test was not considered stringent enough.

Clustering Time Profiles

Principal Component Analysis and Sparse Principal Component Analysis

Multivariate dimension reduction techniques such as PCA (Jolliffe, 2011) and sparse PCA (Huang and Zheng, 2006) can be used to cluster taxa profiles. To do so, we consider as data input the $\boldsymbol{X}$ ($T \times P$) spline fitted matrix. Let $\boldsymbol{t}_1$, $\boldsymbol{t}_2$, …, $\boldsymbol{t}_H$ denote the H principal components of length T and their associated $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, …, $\boldsymbol{v}_H$ factors—or loading vectors, of length P. For a given PCA dimension h, we can extract a set of strongly correlated profiles by considering taxa with the top absolute coefficients in $\boldsymbol{v}_h$. Those profiles are linearly combined to define each component $\boldsymbol{t}_h$, and thus, explain similar information on a given component. Different clusters are therefore obtained on each dimension h of PCA, $h = 1 \ldots H$. Each cluster h is then further separated into two sets of profiles which we denote as "positive" or "negative" based on the sign of the coefficients in the loading vectors (see Results section).

A more formal approach can be used with sparse PCA. Sparse PCA includes ℓ_1 penalizations on the loading vectors to select variables that are keys for defining each component and are highly correlated within a component (see Huang and Zheng, 2006 for more details).

Choice of the Number of Clusters in Principal Component Analysis

We propose to use the average silhouette coefficient (Rousseeuw, 1987) to determine the optimal number of clusters, or dimensions H, in PCA. for a given identified cluster and observation I, the silhouette coefficient of I is defined as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{5}$$

where $a(i)$ is the average distance between observation i and all other observations within the same cluster, and $b(i)$ is the average distance between observation i and all other observations in the nearest cluster. A silhouette score is obtained for each observation and averaged across all silhouette coefficients, ranging from −1 (poor) to 1 (good clustering).

We adapted the silhouette coefficient to choose the number of components or clusters in PCA and sparse PCA (sPCA, i.e., $2 \times H$ clusters), as well as the number of profiles to select for each cluster. Each observation in Eq. (5) now represents a fitted LMMS profile, and the distance between two profiles is calculated using the Spearman correlation coefficient.

Within a given cluster, we calculate the silhouette coefficient of each LMMS profile and apply the following empirical rules for cluster assignation: a coefficient > 0.5 assigns the profile to the cluster, and a value between 0 and 0.5 indicates an uncertain assignment as the profile can be assigned to one or two clusters, while a negative value indicates that the profile should not be assigned to this particular cluster.

To choose the appropriate number of profiles per sPCA component, we perform as follows: for each component, we set a grid of the number of profiles to be retained with sPCA and calculated the average silhouette coefficient per cluster (there are two clusters per component). The final number of profiles to select is arbitrarily set when we observe a sudden decrease in the average silhouette coefficient (see Results section).

Comparison With Functional Principal Component Analysis

Functional principal component analysis (fPCA) has been widely used to cluster longitudinal data by decomposing data matrices into temporal variation models (Hyndman and Ullah, 2007) and has been used in several biological applications (Silverman et al., 1996; Yao et al., 2005). fPCA first models longitudinal profiles into a finite basis of functions then clusters the longitudinal profiles using the basis expansion coefficients of the fPCA scores. fPCA requires the user to choose the number of clusters and the number of components—based on Akaike information criterion, Bayesian information criterion, or percentage of total explained variance, the approach to estimate the fPCA scores—based on conditional expectation or numerical integration, and to cluster the profiles. We used the "fdapace" R package that includes two types of clustering methods, based on model-based clustering of finite mixture Gaussian distribution ("EMCluster") or k-means algorithm based on the fPCA scores.

Evaluation

Clustering

We can assess the quality of clustering with internal measures such as compactness (Dunn, Rand indices, and Jaccard index) or cluster separation. For the latter case, the silhouette coefficient is recognized as an informative criterion Wang et al. (2009) and can be used to compare several clustering results based on the same data. Thus, we used this criterion to assess different methods (PCA, sPCA, and fPCA), or to assess the same method with different parameters—for example, to identify the appropriate number of clusters as we described in 2.4.2. The best clustering approach yields the highest silhouette coefficient.

Measure of Association for Compositional Data

Compositional data arise from any biological measurement made based on relative abundance (Lovell et al., 2015; Gloor et al., 2017). Microbiome data in particular are compositional for several reasons, including biological, technical, and computational. Thus, interpretation based on correlations between profiles must be made with caution as it is highly likely to be spurious. Proportional distances have been proposed as an alternative to measure association. The compositional data analysis field is an active field of research, but methods are critically lacking for longitudinal data. Here, we adopt a practical and *post hoc* approach to evaluate pairwise associations of microbial and omics profiles once they have been assigned to their clusters. We used the proportionality distance φ_s proposed by Lovell et al. (2015) and implemented in the "propr" R package (Quinn et al., 2017). For two LMMS

profiles x_i and x_j, we define the pairwise proportionality distance as

$$\varphi_s\left(x_i,x_j\right)=\frac{\mathrm{var}\left(x_i-x_j\right)}{\mathrm{var}\left(x_i+x_j\right)}. \quad (6)$$

A small value indicates that, in proportion, the pair of profiles is strongly associated. We calculated the distance φ_S on the log-transformed LMMS modeled profiles within each identified cluster to exclude potentially spurious correlations and further guide the interpretation of the results. In addition, to evaluate the quality of our clustering approach, we compared the pairwise distances of the profiles within a particular cluster and profiles outside the cluster.

Integration

Multiblock Projection to Latent Structures Methods

To integrate multiple datasets (also called *blocks*) measured on the same biological samples, we used multivariate methods based on projection to latent structures (PLS) methods (Wold, 1975), which we broadly term *multiblock PLS* approaches. For example, we can consider generalized canonical correlation analysis (GCCA, Tenenhaus and Tenenhaus, 2011; Tenenhaus et al., 2014), which, contrary to what its name suggests, generalizes PLS for the integration of more than two datasets. Recently, we have developed the DIABLO method to discriminate different phenotypic groups in a supervised framework (Singh et al., 2019). In the context of this study, however, we present the sparse GCCA in an unsupervised framework, where input datasets are spline-fitted matrices.

We denote Q data sets $X^{(1)}(TxP_1)$, $X^{(2)}$ (TxP_2), …, $X^{(Q)}(TxP_Q)$ measuring the expression levels of P_q variables of different types (taxa, "omics," continuous response of interest), modeled on T (interpolated) time points, q = 1,…,Q. GCCA solves for each component h = 1,…,H:

$$\max_{a_h^{(1)},\ldots,a_h^{(Q)}} \sum_{q,j=1,q\neq j}^{Q} c_{q,j}\mathrm{cov}\left(X_h^{(q)}a_h^{(q)},X_h^{(j)}a_h^{(j)}\right),$$
$$\text{s.t.}\quad \left\|a_h^{(q)}\right\|_2=1 \text{ and } \left\|a_h^{(q)}\right\|_1\leq\lambda^{(q)} \quad (7)$$

where $\lambda^{(q)}$ is the ℓ_1 penalization parameter, $a_h^{(q)}$ is the loading vector on component h associated with the residual (deflated) matrix $X_h^{(q)}$ of the data set $X^{(q)}$, and $C = \{c_{qj}\}$ is the design matrix. C is a $Q\times Q$ matrix that specifies whether datasets should be correlated and includes values between zero (datasets are not connected) and one (datasets are fully connected). Thus, we can choose to take into account specific pairwise covariances by setting the design matrix (see Rohart et al., 2017 for implementation and usage) and model a particular association between pairs of datasets, as expected from prior biological knowledge or experimental design. In our integrative case study, we used sparse PLS, a special case of Eq. (7) to integrate microbiome and metabolomic data, as well as sparse multiblock PLS to also integrate variables of interest. Both methods were used with a fully connected design.

The multiblock sparse PLS method was implemented in the `mixOmics` R package where the ℓ_1 penalization parameter is replaced by the number of variables to select, using a soft-thresholding approach (see more details in Rohart et al., 2017).

Parameter Tuning

The integrative methods require choosing the number of components H, defined as $t_h^{(q)}=X_h^{(q)}a_h^{(q)}$, and number of profiles to select on each PLS component and in each dataset. We generalized the GCCA approach by using the silhouette coefficient based on a grid of parameters for each dataset and each component.

Simulation and Case Studies

Simulation Study Description

A simulation study was conducted to evaluate the clustering performance of multivariate projection-based methods such as PCA, and the ability to interpolate time points in LMMS.

Twenty reference time profiles were generated on nine equally spaced time points and assigned to four clusters (five profiles each). These ground truth profiles were then used to simulate new profiles. We generated 500 simulated datasets.

Clustering Performance

We first compared profiles simulated then modeled with or without LMMS:

A) For each of the reference profiles, five new profiles (corresponding to five individuals) were sampled to reflect some inter-individual variability as follows: let x be the observation vector for a reference profile r, r = 1…20; for each time point t (t = 1,…,9), five measurements were randomly simulated from a Gaussian distribution with parameters $\mu=x_{t,r}$ and σ^2, where σ ={0,0.1,0.2,0.3,0.4,0.5,1,1.5,2,3} to vary the level of noise. This noise level was representative of the data described below. The profiles from the five individuals were then modeled with LMMS, resulting in 500 matrices of size (9×20) for each level of noise σ.

B) For each of the reference profiles, one new profile was simulated as described in step A, but no LMMS modeling step was performed, resulting in 500 matrices of size (9×20) for each level of noise σ.

Clustering was obtained with PCA and compared to the reference cluster assignments in a confusion matrix. The clustering was evaluated by calculating the accuracy of assignment $(\frac{TP+TN}{TP+FP+TN+FN})$ from the confusion matrix, where for a given cluster, TP (true positive) is the number of profiles correctly assigned in the cluster, FN (false negative) is the number of profiles that have been wrongly assigned to another cluster, TN (true negative) is the number of profiles correctly assigned to another cluster, and FP (false positive) is the number

of profiles incorrectly assigned to this cluster. Besides accuracy, we also calculated the Rand index (Rand, 1971) objective as a similarity metric to the clustering performance of PCA. The clustering results from fPCA were poor, even for a low level of noise (**Supplementary Figure 1**); thus, fPCA was not compared against PCA.

Interpolation of Missing Time Points

To evaluate the ability of LMMS to predict the value of a missing time point for a given feature over time, we randomly removed 0 to 4 measurement points in the simulated datasets described above in step A. We compared the PCA clustering performance with or without LMMS interpolation.

Infant Gut Microbiota Development

The gastrointestinal microbiome of 14 babies during the first year of life was studied by Palmer et al. (2007). The authors collected an average of 26 stool samples from healthy full-term infants. As infants quickly reach an adult-like microbiota composition, we focused our analyses on the first 100 days of life. Infants who received an antibiotic treatment during that period were removed from the analysis, as antibiotics can drastically alter microbiome composition (Dudek-Wicher et al., 2018).

The dataset we analyzed included 21 time points on average for 11 selected infants (vaginal delivery = 6, C-section = 5; see **Figure 3**). Samples were collected daily during days 0–14 and weekly after the second week. We separated our analyses based on the delivery mode (C-section or vaginal), as this is known to have a strong impact on gut microbiota colonization patterns and diversity in early life Rutayisire et al. (2016). The purpose of our statistical analysis was to identify a bacterial signature that describes the dynamics of a baby's microbial gut development in the first days of life, as well as compare differences in signatures between babies born by vaginal delivery or by C-section. As this study is single omics, we applied our framework depicted in **Figure 2** with sPCA.

Waste Degradation Study

Anaerobic digestion (AD) is a highly relevant microbial process to convert waste into valuable biogas. It involves a complex microbiome that is responsible for the progressive degradation of molecules into methane and carbon dioxide. In this study, AD's biowaste was monitored across time (more than 150 days) in three lab-scale bioreactors as described in (Poirier et al., 2016).

We focused our analysis on days 9 to 57, which correspond to the most intense biogas production. Degradation performance was monitored through four parameters: methane and carbon dioxide production (16 time points) and the accumulation of

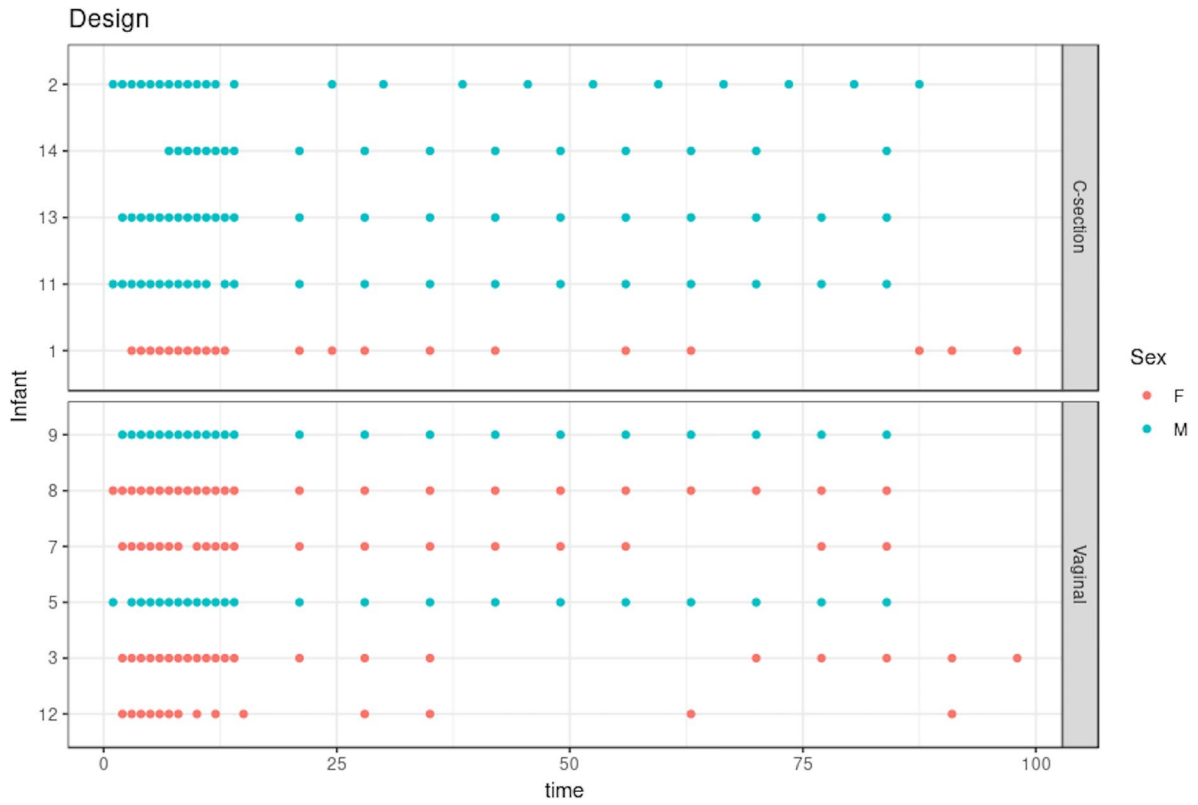


FIGURE 3 | Infant gut microbiota development study: stool samples were collected from six male and five female babies over the course of 100 days. Samples were collected daily during days 0–14 and weekly thereon until day 100. Time is indicated on the x-axis in days. As delivery method is known to be a strong influence on gut microbiome colonization, the data are separated according to either C-section or vaginal birth.

acetic and propionic acid in the bioreactors (5 time points). Microbial dynamics were profiled with 16S RNA gene metabarcoding as described in Poirier et al. (2016) and included 4 time points and 90 OTUs. A metabolomic assay was conducted on the same biological samples at four time points with gas chromatography coupled to mass spectrometry GC-MS after solid phase extraction to monitor substrates degradation (Limam et al. (2010). The `XCMS` R package (version 1.52.0) was used to process the raw metabolomics data (Smith et al., 2006). GC-MS analyses focused on 20 peaks of interest identified by the National Institute of Standards and Technology database. Data were then log-transformed. The purpose of the study was to investigate the relationship between biowaste degradation performance and microbial and metabolomic dynamics across time. The aim of our statistical analysis was to identify highly associated multi-omic signatures characterizing waste degradation dynamics in the three bioreactors. This study involves the integration of two omics datasets and degradation performance measures; thus, we applied sPLS and multiblock sPLS, as shown in our workflow in **Figure 2**.

RESULTS

Simulation Study

Clustering Performance

Figure 4 shows the clustering performance of PCA with an increasing amount of noise in the simulated profiles. Unsurprisingly, PCA gave optimal clustering performance when noise was absent, with or without profile modeling to take into account individual variability. When noise increased, PCA performed better with modeling, which acts as a denoising process. Finally, a high level of noise showed the limitation of the modeling approach, as similar clustering results were obtained with or without LMMS modeling. However, the PCA clustering performance was still very good, with a mean accuracy of 0.7 when the level of noise was maximum.

Interpolation of Missing Time Points

We evaluated the ability of LMMS to interpolate an increasing number of missing time points (up to four). Interpolation is important in our framework as it allows the estimation of evenly spaced time points as well as time points that may be missing in

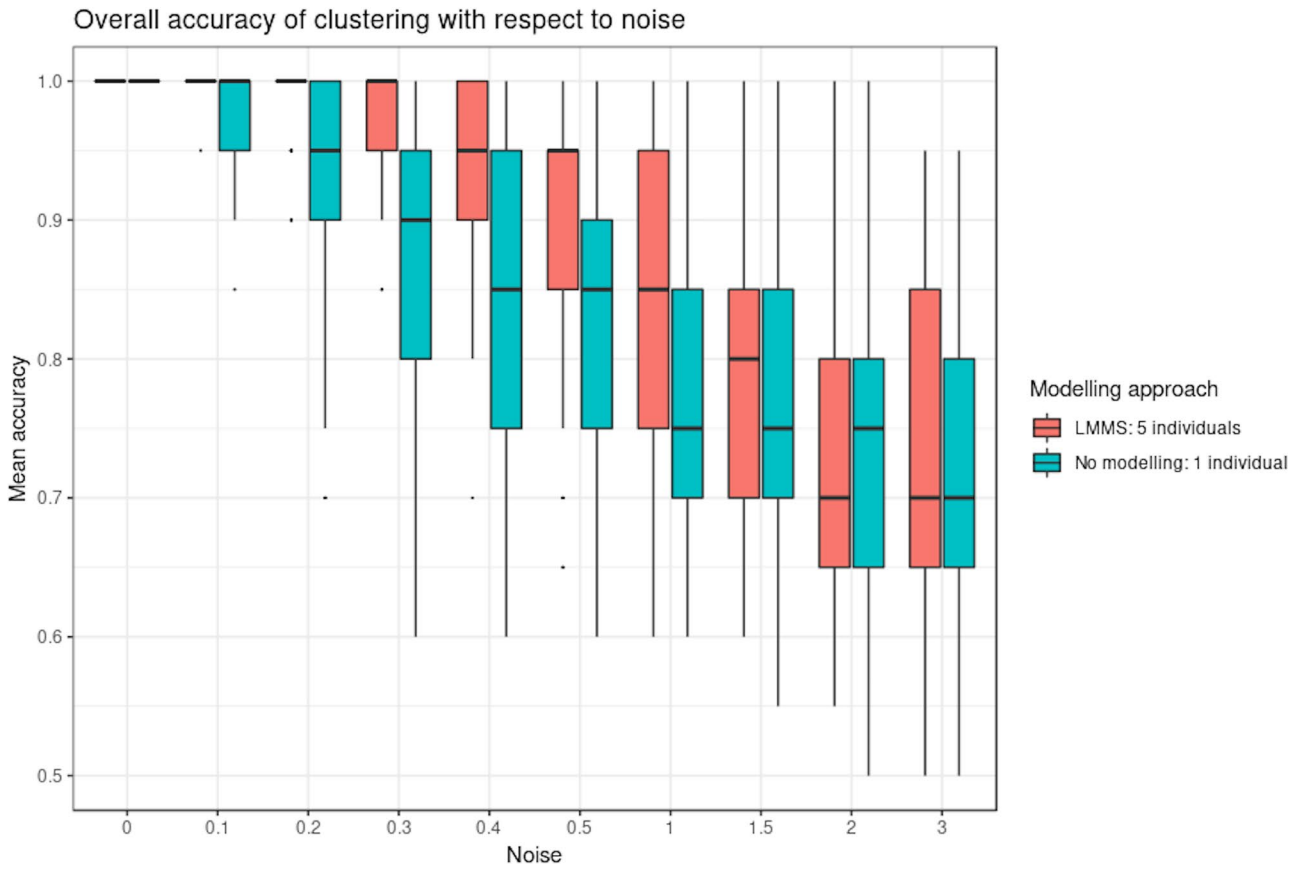


FIGURE 4 | Simulation study: overall accuracy of clustering with respect to noise. Twenty reference profiles, which grouped into four clusters were used as a basis for simulation, and each of the new simulated profiles were generated with random noise. We compared two approaches: with linear mixed model spline (LMMS) modeling: five new profiles were generated per reference, and without modeling: only one profile was simulated per reference. We evaluated the ability of principal component analysis clustering to correctly assign the simulated profiles in their respective reference clusters based on mean accuracy: without noise, both approaches lead to a perfect clustering; with noise < 1, LMMS modeling acts as a denoising process with better performance than no modeling; and with a high level of noise ≥ 1, the performances of both approaches decrease.

one data set but not in the other (e.g., biowaste degradation study). Interpolation did not seem to affect the clustering performance of PCA (**Figure 5** and **Supplementary Figure 2**). Rather, the level of noise had the largest impact on clustering: the mean accuracy was close to 1 when the noise was nonexistent but decreased as the number of missing time points and noise increased. In the latter scenarios, LMMS interpolation seemed to give, on average, better clustering than without interpolation. When the number of missing time points increased, we observed a better classification accuracy with noise compared to no noise. This can be explained by the LMMS modeling of straight lines in the latter case that led to poor clustering (**Supplementary Figure 3**).

Clustering Time Profiles: Infant Gut Microbiota Development Study

Pre-Processing and Modeling

A total of 2,149 taxa were identified in the raw data (**Table 1**). After the pre-processing steps illustrated in **Figure 2**, a smaller number of OTUs were found in fecal samples of babies born by C-section than vaginal delivery. Similarly, a simple linear regression model showed a smaller proportion of OTUs in babies born *via* C-section (73%) than vaginal delivery (81%), and this was also observed after the filtering step (**Table 1**).

TABLE 1 | Infant gut microbiota development study: number of operational taxonomic units (OTUs) identified and linear model types fitted according to delivery mode.

		C-section	Vaginal
Identified OTUs		2,149	2,149
Number of OTUs after pre-processing		107	117
Linear model types	(1)	78	95
	(2)	29	22
Linear model types after filtering	(1)	42	68
	(2)	29	22

Comparison of Principal Component Analysis and Functional Principal Component Analysis

According to our tuning criteria, we obtained four clusters with PCA (i.e., two components). We therefore set the same number of clusters in fPCA for comparative purposes. PCA clustering outperformed fPCA for each delivery mode dataset that was analyzed (see **Table 2**). The resulting fPCA clustering is displayed in **Figure 6** for babies born *via* vaginal delivery. We found that the EM approach in fPCA tended to cluster a larger number of uncorrelated OTUs compared to the *k*-CFC approach (average silhouette coefficient = 0.07 for EM and 0.61 for *k*-CFC).

We used sPCA to select key OTU profiles for each cluster. This step is essential for discarding profiles that are distant from the

TABLE 2 | Infant gut microbiota development study: average silhouette coefficient according to clustering method.

	PCA	sPCA	fPCA (*k*-CFC)	fPCA(EM)
Vaginal	0.84	0.95	0.61	0.07
C-section	0.87	0.86	0.69	0.35

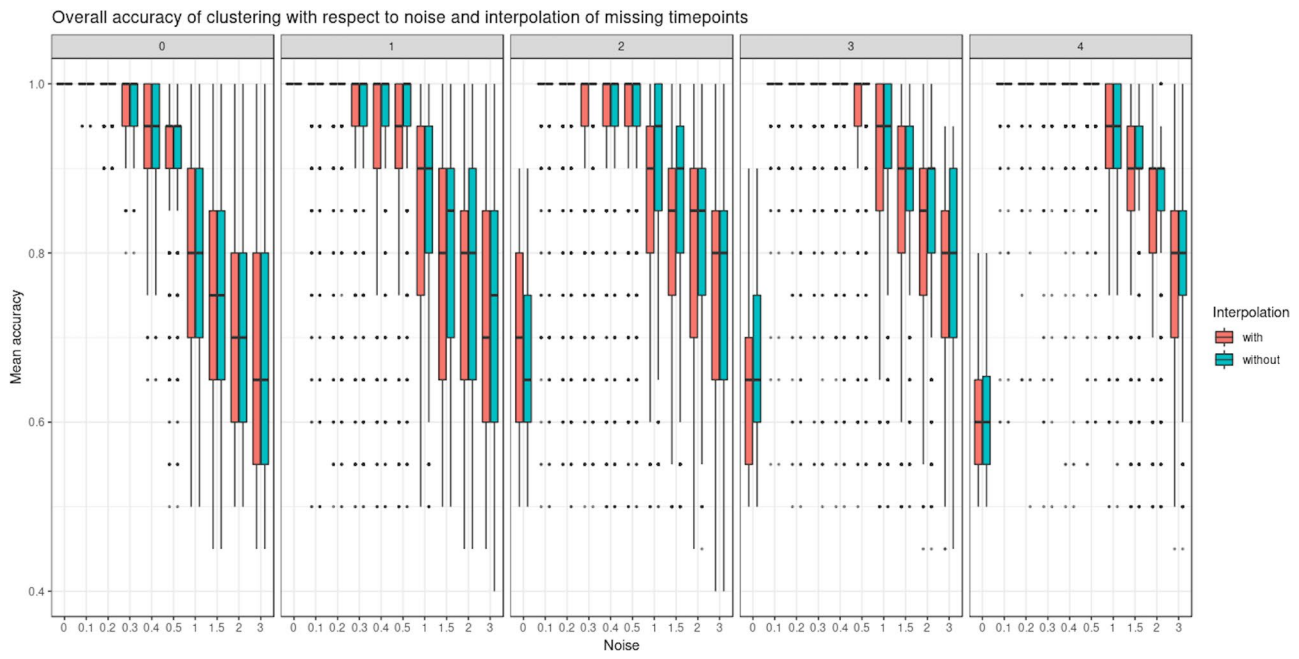


FIGURE 5 | Simulated study: overall accuracy of clustering when time points are missing. The simulation scheme is described in 2.7.1; however, here, some time points were removed. We compared the ability of linear mixed model spline (LMMS) to interpolate missing time points. When there are no time points missing, both interpolated and non-interpolated approaches gave a similar performance. When the number of time points increases, the classification accuracy decreases. Without noise and with several time points removed, LMMS tended to model straight lines, resulting in poor clustering (see also **Supplementary Figure 3**).

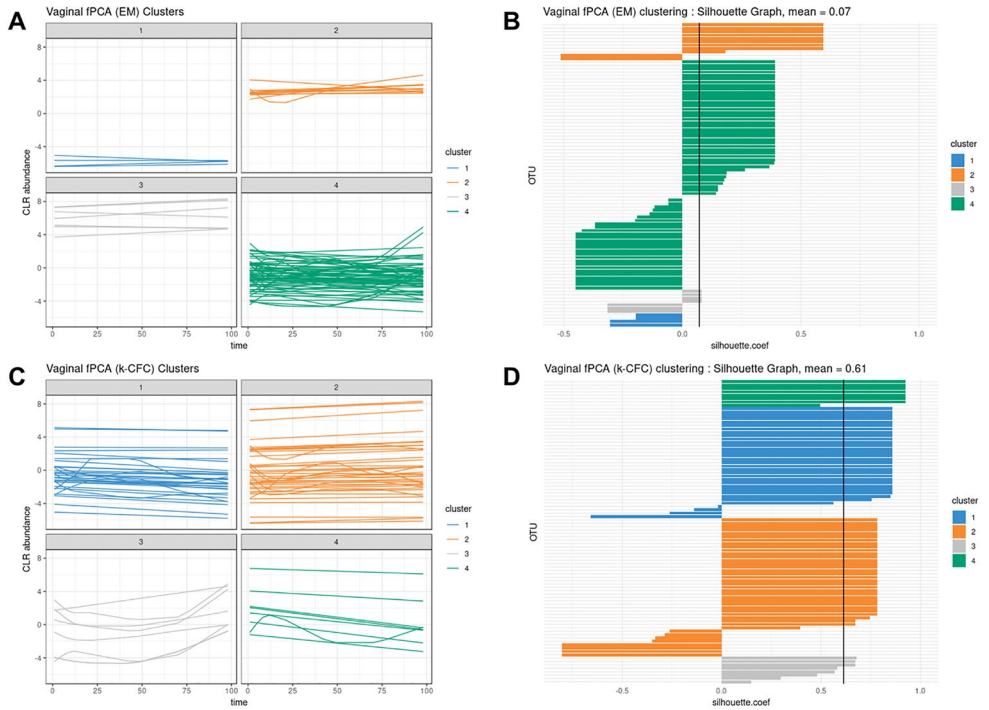


FIGURE 6 | Infant gut microbiota development. Functional principal component analysis expectation–maximization clustering (first row) and *k*-center functional clustering (second row). **(A–C)** Vaginal operational taxonomic unit (OTU) profiles clustered with either EM or *k*-CFC. Each line represents the relative abundance of a selected OTU across time. **(B–D)** Silhouette coefficients for each profile and each clustering. Each bar represents the silhouette coefficient of a particular OTU, and colors represent assigned clusters. The average coefficient is represented by a vertical black line. The average silhouette coefficient was 0.07 for EM clustering and 0.61 for *k*-CFC clustering.

average cluster profile and thus not informative. As expected, we observed an overall increase in the silhouette average coefficient for the sPCA clustering compared to PCA, indicating a better clustering capability (see **Table 2**). According to the silhouette average coefficient, vaginal delivery showed the best partitioning for PCA clustering (0.87; **Table 2**). Cluster 1 (denoted "component 1 positive" in **Figure 7A**) showed a relative increase in abundance of species, including some that are characteristic of a healthy "adult-like" gut microbiome composition such as the clade *Bacteroidetes* (Thursby and Juge, 2017). The proportionality distance within cluster 1 was low (**Supplementary Table 1**), with a strong association between *Bacteroides* and *Fusobacteria* (φ_s = 0.04), as well as between *Actinobacter* with *Bacteroides* (φ_s = 0.02) and *Fusobacteria* (φ_s = 0.09). According to this distance, there might have been a spurious correlation identified between the genus *Bacteroides* and an environmental uncultured bacterium *(clone HuCA36)* (φ_s = 14.81); see **Supplementary Table 2**. In cluster 2 ("component 1 negative"), relative profile abundance tended to decrease and corresponded to genera found in vaginal and skin microbiota, such as *Lactobacillus* and *Propionibacterium* (Grice and Segre, 2011; Bing et al., 2012). According to the proportionality distance, *Propionibacterium* and *Lactobacillus* were highly associated (φ_s = 0.29) as well as with *Campylobacter* (φ_s = 0.39, see **Supplementary Table 2**). Clusters 3 and 4 (denoted "component 2 positive and negative") highlighted taxa profiles with negative association.

A cladogram representing all OTUs and those selected by sPCA for each cluster is shown in **Figure 8** and illustrates that most families are presented in our OTU selection. In addition, we can observe specific clusters—family patterns as discussed above.

Thus, with this preliminary PCA analysis, we were able to rebuild a partial history behind the development of the gut microbiota. Vaginal species that initially colonized in the gut progressively disappeared to enable species that characterize adult gut microbiota.

For babies born by C-section, four clusters were identified by PCA (**Figure 7D**; cladogram visualization is available in **Supplementary Figure 4**). The median values of the proportionality distance within the different clusters were significantly lower than between the selected OTUs in the clusters and all the other OTUs (**Supplementary Table 3**). For example, the median value within cluster 1 was 0.11 compared to 1.36 outside the cluster. Clusters 1 and 2 ("component 1 positive and negative") displayed either an increase or decrease in relative abundance. However, none of the cluster 2 species are known to characterize, or were found in, vaginal delivery, suggesting that the infant gut was first colonized by the operating room microbes as already demonstrated by Shin et al. (2015). Cluster 3 ("component 2 positive") revealed transitory states of increase then decrease of relative abundance profiles, while cluster 4 ("component 2 negative") showed the reverse trend.

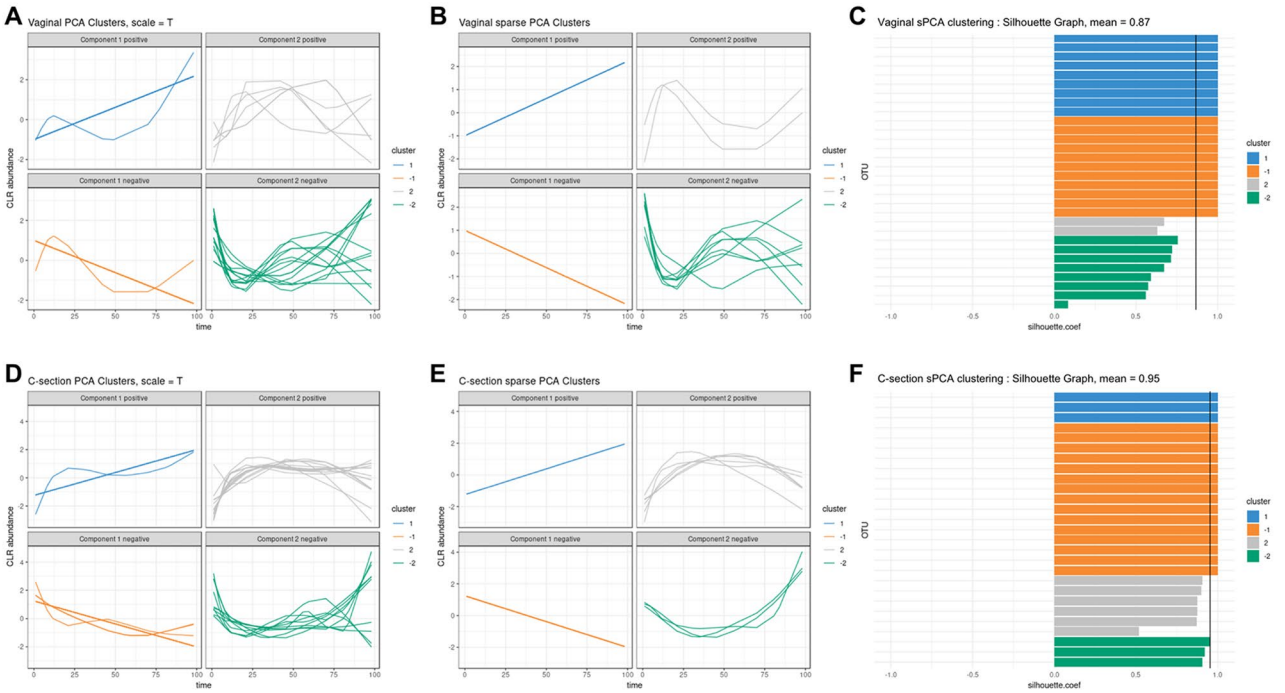


FIGURE 7 | Infant gut microbiota development. Vaginal (first row) and C-section delivery babies (second row). **(A**, **C)**, **(D**, **E)** operational taxonomic unit (OTU) profiles clustered with either principal component analysis (PCA) or sparse PCA (sPCA). Each line represents the abundance of an OTU across time. OTUs were clustered according to their contribution on each component for PCA or sPCA (that includes variable selection). The PCA clusters were further separated into profiles denoted "positive" or "negative" that refer to the sign of the loading vector from s/PCA. Profiles were scaled to improve visualization. **(C**–**F)** Silhouette profiles for each identified clustering. Each bar represents the silhouette coefficient of a particular OTU, and colors represent assigned clusters. The average coefficient is represented by a vertical black line. A greater average silhouette coefficient means a better partitioning state. **(C)** Vaginal delivery babies with sPCA (average silhouette coefficient = 0.87). **(F)** C-section delivery babies with sPCA (average silhouette coefficient = 0.95).

When comparing the dynamics of the two delivery methods, we found a higher diversity in the intestinal microbiota of babies born vaginally (117 modeled profiles) than by C-section (107). For vaginal delivery, the modeling step identified a larger proportion of straight lines, which may indicated a greater inter-individual variability compared to C-section delivery. The clusters denoted "component 1 positive" in both delivery modes showed an increased relative abundance over time, with 32 OTUs assigned to this cluster in vaginally born babies, compared to 11 in C-section (**Table 3**). Despite the relatively sterile environment of the operating room, it was surprising to observe similar number of OTUs in cluster "component 1 negative" for both types of delivery mode (vaginal: 38, C-section: 35), as we would have expected to identify a larger number of opportunistic microorganisms colonizing babies born vaginally(e.g., *Propionibacterium acnes, Campylobacter*). These include species found on the surface of the skin and in the vaginal flora. However, for babies born by C-section, we observed a large number of microorganisms from various origins (e.g., *Staphylococcus*, *Rickettsia*, *Rhodobacter*).

In summary, sparse PCA clustering of LMMS modeled profiles enabled the identification of groups of microorganisms with relative increased abundance over time. These microorganisms are characteristics of an adult gut microbiota. We also identified groups of opportunistic microorganisms with a decreasing relative abundance over time. We also found that, during the first year of life, gut microbiota was more diverse for babies born by vaginal than C-section delivery.

Clustering Omics: Waste Degradation Study

Pre-Processing and Modeling

A total of 90 OTUs were identified in the 12 samples of the initial dataset (**Table 4**). After pre-processing, 51 OTUs were retained. Approximately 60% (resp. 50%) of the OTUs (resp. metabolites) were fitted with linear regression models (1), and 40% (resp. 50%) were modeled by more complex spline models (2)–(4). All performance measures were also modeled by splines. During the filtering step, seven OTUs and four metabolites that were fitted with linear regression models were discarded. The small number of profiles that were filtered out indicated that the variability between the three bioreactors was relatively low.

Sparse PCA on Concatenated Datasets

As a first and naive attempt to jointly analyze microbial, metabolomic, and performance measures, all three datasets were concatenated

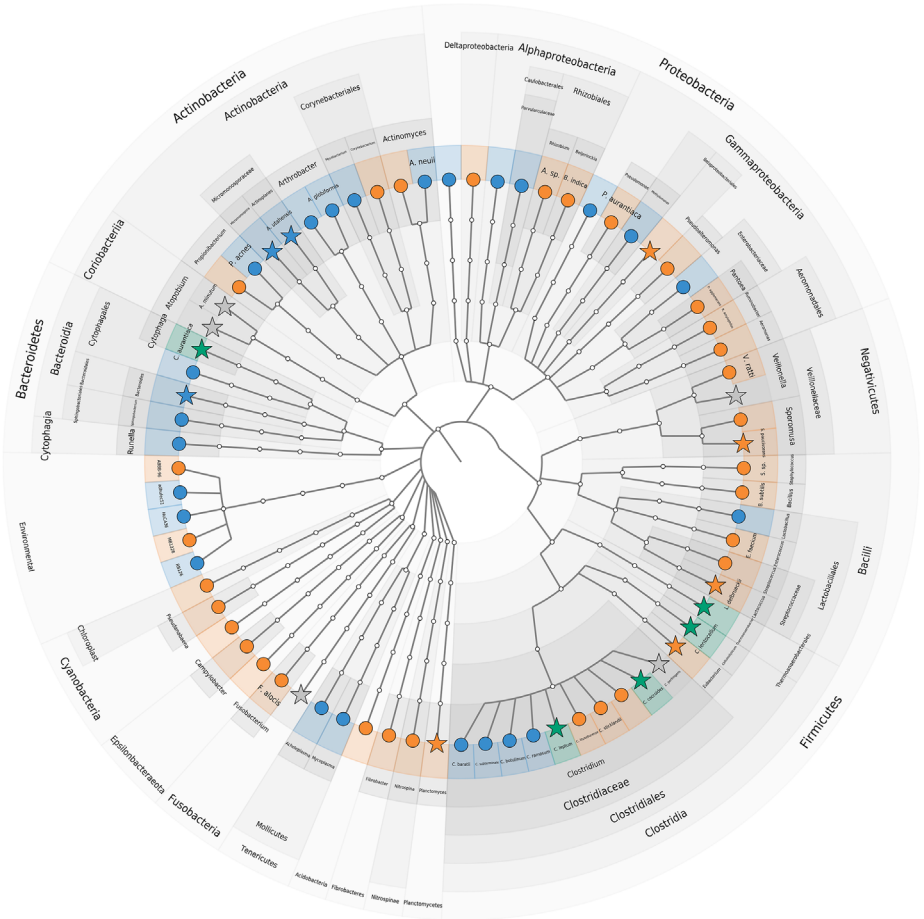


FIGURE 8 | Infant gut microbiota development study, vaginal data. Cladogram representing the principal component analysis (PCA) clustering results using GraphlAn (Asnicar et al., 2015). Terminal nodes are colored according to the clustering, and operational taxonomic units selected with sparse PCA are represented with a star.

then analyzed with sPCA. Only a very small number of profiles from the different datasets were selected. This small selection is likely due to the high variability in each data type. Selected variables included mainly OTUs and performance measures. These were assigned to four clusters and included respectively 1, 3, 2, and 3 OTUs with 0, 1, 2, and 0 metabolites and 2, 0, 1, and 0 performance measures. The average silhouette coefficient was 0.744, a potentially sub-optimal clustering compared to our analyses presented in the next section. This preliminary investigation highlighted the limitation of sPCA to identify a sufficient number of associated profiles from disparate sources.

Microbiome-Metabolomic Integration With sPLS

The results from the sPLS analysis are shown in **Supplementary Figure 5**. Four clusters of variables were identified, and the average silhouette coefficient of 0.954 confirmed that sPLS led to better clustering of the different types of profiles than sPCA. The proportionality distances of the profiles within each cluster are presented in **Table 5** and in **Supplementary Figure 6**. Their low values indicated strong associations between profiles within each cluster, compared to any association outside each of the clusters. A cladogram representing the selected OTUs only, according to each sPLS cluster is shown in **Supplementary Figure 7**.

TABLE 3 | Infant gut microbiota development study: number of operational taxonomic units (OTUs) per cluster identified with principal component analysis (PCA) clustering and OTUs selected in brackets with sparse PCA.

	C-section	Vaginal
Cluster 1 (comp 1 positive)	11 (3)	32 (9)
Cluster 2 (comp 1 negative)	35 (15)	38 (11)
Cluster 3 (comp 2 positive)	15 (6)	6 (2)
Cluster 4 (comp 2 negative)	10 (3)	14 (8)

The first cluster (denoted "component 1 negative") included 10 OTUs and 4 metabolite variables and showed increasing

TABLE 4 | Waste degradation study: operational taxonomic units (OTUs), metabolites, and performance modeling and filtering in the bioreactor study. Only OTU data were pre-processed.

Type of features		OTUs	Metabolites	Performance
Number of features		90	20	4
Number of Features after pre-processing		51	NA	NA
Linear model types	(1)	30	10	0
	(2)	19	0	2
	(3)	2	4	0
	(4)	0	6	2
Linear model types after filtering	(1)	24	6	0
	(2)	19	0	2
	(3)	2	4	0
	(4)	0	6	2

TABLE 5 | Waste degradation study: proportionality distance for clusters identified with sparse PLS. The median distance between all pairs of profiles, within cluster, and with the entire background set (outside a given cluster) is reported. A Wilcoxon test p-value assesses the difference between the medians.

Cluster	Median within cluster	Median outside cluster	Wilcoxon test P-value
1 (comp 1 positive)	0.43	1.37	$9.40*10^{-57}$
–1 (comp 1 negative)	0.42	1.11	$1.76e*10^{-28}$
2 (comp 2 positive)	0.29	0.97	$5.71*10^{-24}$
–2 (comp 2 negative)	0.01	0.87	$2.82*10^{-13}$

relative abundance until a plateau was reached at approximately 40 days. Median value of the proportionality distance within the cluster was 0.42, which was compared to 1.11 between the variables selected in the cluster and all the other variables, indicating strong associations within this cluster. The OTUs were microorganisms often recovered during AD of biowaste, such as methanogenic archaea of *Methanosarcina* genus or bacteria of *Clostridiales*, *Acholeplasmatales*, and *Anaerolineales* orders. These were reported as being involved in the different steps of AD (Poirier et al., 2016). Their relative abundance increased while biowaste was degraded, until there was no more biowaste available in the bioreactor.

From the proportionality distances, we found that their abundance across time was, in proportion, similar, indicating a synchronized role during this biological process. In particular, of all the proportionality distances between the profiles of archaea of *Methanosarcina* genus and bacteria of *Clostridiales* order, the *Syntrophomonadaceae* family was the lowest which made sense as these microorganisms have already been reported as syntrophs (Liu et al., 2011); see **Supplementary Table 4**.

Their abundance was also highly associated, in proportion, to the intensity of various metabolites produced during the AD process, such as benzoic acid that is formed during the degradation of phenolic compounds (Hoyos-Hernandez et al., 2014), or phytanic acid, known to be produced during the fermentation of plant materials in the ruminant gut (Watkins et al., 2010), as well as indole-2-carboxylic acid. Thus, the identified microorganisms were likely responsible for the production of these compounds. Cluster 2 (component 1 positive) included 10 OTUs and 4 metabolites. The median value of the proportionality distance within the cluster was also very low compared to the proportionality distance outside the cluster (0.29 and 0.97; **Table 5**). Profiles of cluster 2 were negatively correlated to cluster 1, and their relative abundance decreased with time. OTUs mainly belonged to the *Bacteroidales* order. They were present in the initial inoculum but did not survive in this experiment, as the operating conditions or the substrate were not optimal for their growth, as observed in other studies (Madigou et al., 2019). Consequently, their relative abundance progressively decreased over time. Metabolites identified in cluster 2 were present in the biowaste and were degraded during the experiment. They included fatty acids (decanoic and tetradecanoic acids) that can be found in oil, or 3-(3-hydroxyphenyl)propionic acid, arising from the digestion of aromatic amino acids or breakdown product of lignin or other plant-derived phenylpropanoids. As their profile was negatively correlated to those from cluster 1, it is likely that these metabolites were consumed by OTUs assigned to cluster 1 (Torres et al., 2003). Cluster 3 (component 2 negative) included one OTU and five metabolites. Profiles relative abundance decreased slowly with time until reaching a stable abundance after 20 days. One OTU of *Clostridiales* order appeared to have been out-competed by other OTUs or phase active only during the first days of the degradation, which corresponds to the degradation of complex biopolymers contained in biowaste (Poirier et al., 2016). Among the metabolites of this cluster, hydrocinnamic and 3,4-dihydroxyhydrocinnamic acids are commonly found in plant biomass and its residues (Boerjan et al., 2003). Their molecular structure may have contributed to their slower degradation compared to other molecules, which may explain their stable abundance in the digesters until day 30. Finally, cluster 4 (component 2 positive) included 11 OTUs and 3 metabolites with slow relative abundance increase. OTUs from this group were very varied with eight orders represented. They may have had slower growth rates than OTUs of cluster 1 or were possibly involved in the degradation of molecules from cluster 3. Their abundance may also have had a slow increase as they fed on specific molecules that are only formed during the digestion process. Metabolites included N-acetylanthranilic acid and dehydroabietic acid that were likely produced by microorganisms and accumulated during the AD process, suggesting they could not be metabolized by other microorganisms.

Integration of Microbiome, Metabolomic and Performance Data with MultiBlock sPLS

Figure 9 illustrates the results from the integration of the three datasets, where the performance data are considered as the response of interest. Similar to the sPLS analysis, block sPLS assigned profiles to four clusters, with an average silhouette coefficient of 0.909. The proportionality distances are summarized in **Figure 10** and in **Supplementary Table 5** and show a greater level of association between profiles within each cluster, compared to the associations with all other profiles outside the cluster (see **Supplementary Figure 8** per omic variable).

Two performance variables (methane and carbon dioxide productions) were assigned to cluster 1 (component 1 negative). This result is biologically relevant, as biogas is the final output of

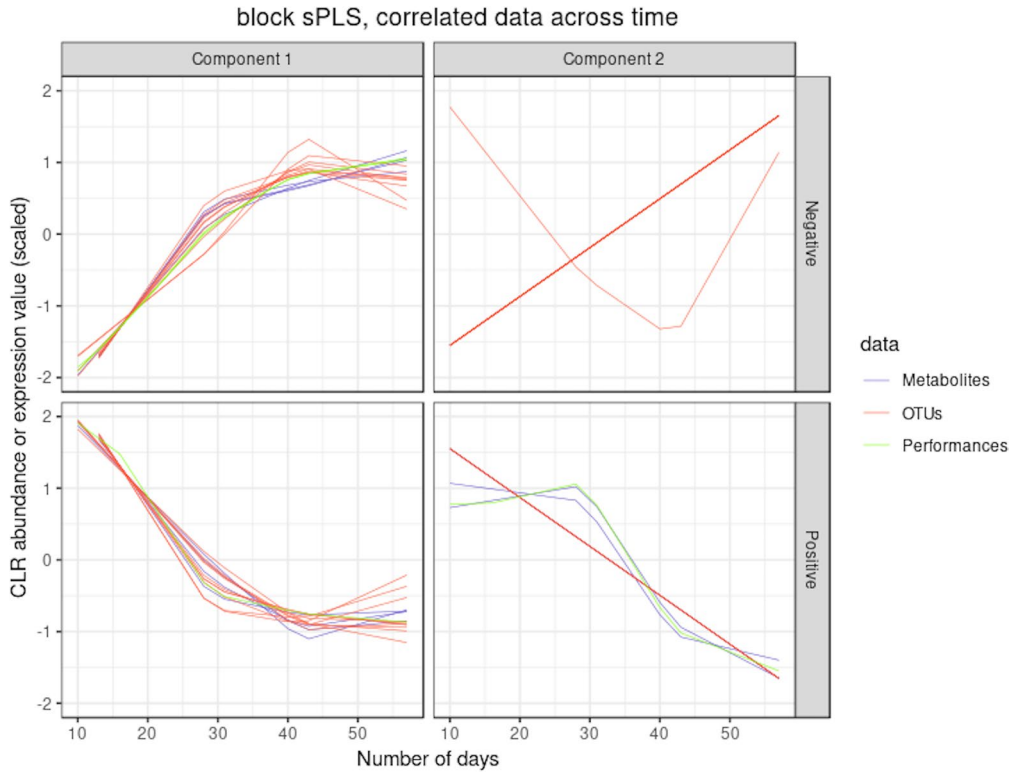


FIGURE 9 | Waste degradation study: integration of OTUs, metabolites and performance measures with multiblock sPLS. Each line represents the relative abundance of OTUs, metabolites and performance measures selected by multiblock sPLS across time. OTUs, metabolites and performance measures were clustered according to their contribution on each component. The clusters were further separated into profiles denoted `positive' or `negative' that refer to the sign of the loading vector from multiblock sPLS.

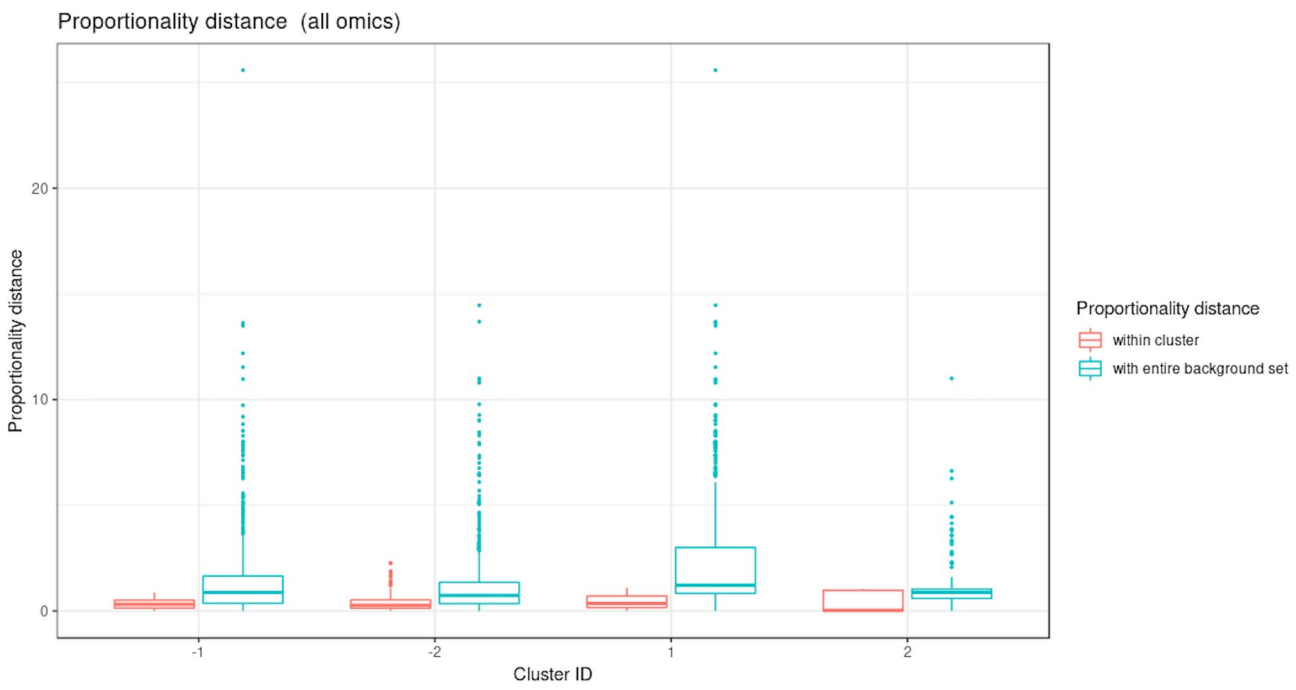


FIGURE 10 | Waste degradation study: Proportionality distance per cluster identified with multiblock sparse PLS. The distance was calculated between each pair of profiles within a given cluster and with the entire background set (outside a given cluster), for all omics.

the AD reaction and is known to be associated with microbial activity and growth. Moreover, it is produced by archaea, such as *Methanosarcina*, which is also selected in this cluster. The proportionality distance between this OTU and methane was very low (φs = 0.25; **Supplementary Table 4**) confirming a strong association. Cluster 1 therefore represented the progress of the degradation process. In Cluster 2 (component 1 positive), we identified acetate produced by bacteria in the early days of the incubation and consumed by archaea (cluster 1) to produce biogas. It was logically negatively associated to cluster 1 representing the progress of the degradation. Propionate was assigned to cluster 3 (component 2 positive). Its degradation was delayed compared to the molecule of cluster 1. It was expected as, for thermodynamical reasons, its degradation usually only starts when all acetate is degraded (Chapleur et al., 2014). It was biologically relevant to find it associated with hydrocinnamic and 3,4-dihydroxyhydrocinnamic acids, which are also difficult to degrade. Cluster 4 (component 2 negative) was composed of only OTUs and metabolites and was similar to the one obtained with sPLS on component 2 positive.

In summary, our framework allowed us to integrate different omic datasets measured longitudinally and identify subsets of relevant microorganisms that were highly associated with metabolites abundance and performance measures through the biodegradation process. These analyses constitute a first step toward generating novel hypotheses about the biological mechanisms underpinning the dynamics in AD.

DISCUSSION

Advances in technology and reduced sequencing costs have resulted in the emergence of new and more complex experimental designs that combine multiple omic datasets and several sampling times from the same biological material. Thus, the challenge is to integrate longitudinal, multi-omic data to capture the complex interactions between these omic layers and obtain a holistic view of biological systems. In order to integrate longitudinal data from microbial communities with other omics, meta-omics, or other clinical variables, we proposed a data-driven analytical framework to identify highly associated temporal profiles between these multiple and heterogeneous datasets.

The application of this method allows the identification of similar expression profiles within a particular dataset (e.g., infant gut microbiota development study) but also across heterogeneous data types (16S amplicon microbiome data, metabolomics, chemical data in the waste degradation study). The clustering of longitudinal profiles helps identify groups of biological entities that may be functionally related and thus generate novel hypotheses about the regulatory mechanisms that take place within the ecosystem.

In the proposed framework, the microbial counts of the microbiota's constituent species are normalized for uneven sequencing library sizes and compositional data. Modeling with linear mixed model splines enables us to reduce the dimension of the data across the different biological replicates and take into account the individual variability due to either technical or biological sources. This approach also enables us to compare data analyzed at different time points (e.g., the waste degradation study). Lastly, we clustered the data using multivariate dimension reduction techniques on the spline models that further allowed integration between different data types, and the identification of the main patterns of longitudinal variation.

Ribicic et al. (2018) proposed an approach similar to ours, but they applied individual PCA or sPCA on each dataset (chemical loss and microbial community) after local polynomial regression modeling. Integration was performed in a second stage of the analysis with PLS by using hierarchical clustering (Cluster Image Maps visualization) to identify correlations between the two datasets. In comparison, we offer a more complete framework that accommodates complex scenarios, across several omics and across replicates, and handles compositional data. The LMMS allows for the modeling of expression over time for each compound across biological replicates while taking into account the overall individual variability. We used sPCA, sPLS, and block sPLS as clustering means by leveraging on the loading vectors from these methods while selecting meaningful profile signatures.

Integrating different types of microbiome longitudinal data (e.g., abundance, activity, metabolic pathways, or macroscopic output) can be naively performed by concatenating all datasets. However, we showed that this approach was unsuccessful at selecting a sufficiently large number of profiles of different types and thus did not shed light on the holistic view of the ecosystem dynamics (bioreactor study). Our integrative multivariate methods sPLS and block sPLS were better suited for the integration task, as they do not merge but rather statistically correlate components built on each dataset, and thus avoid unbalance in the signature when one dataset is either more informative, less noisy, or larger than the other datasets.

When compared with fPCA, which uses either *k*-CFC or EM clustering algorithms, we showed that our approach led to better clustering performance. In addition, the sparse multivariate approaches sPCA and block sPLS enabled the identification of key profiles to improve biological interpretation. Note however that fPCA might be better suited than our approach for a large number of time points, as we discuss next.

We have identified several limitations in our proposed framework. First, a high individual variability between biological replicates limits the LMMS modeling step, resulting in simple linear regression models to fit the data. While a straight line model may accurately describe temporal dynamics, it could also be due to a poor quality of fit. We have implemented the Breusch–Pagan test to address this issue. Alternatively, in the case of a very high inter-individual variability that prevents appropriate smoothing, one could consider *N of One* analyses as proposed by (Gerber et al. (2012); Äijö et al. (2017) with time dynamical probabilistic models.

Second, a large number of time points can result in the modeling of noisy profiles and clusters, often due to high individual variability. Highly variable and vastly different profiles can also be difficult to cluster appropriately. Therefore, this framework is recommended when the number of time points remains small (5–10) and when regular and similar trends are expected from the data.

Third, even though our simulation results showed that the LMMS interpolation of missing time points did not seem to

impact clustering, the overall performance of the approach would be optimal for regularly spaced time points in the omics longitudinal experiments.

Fourth, we have not fully addressed the issue of analyzing time-course compositional data. Indeed, when working with relative abundances, fluctuations in the abundance of a particular microorganism might result in spurious fluctuations in the abundance of other microorganisms. This issue is not specific to microbiome data only, as other sequenced-based data are intrinsically compositional (Gloor et al., 2017). Thus, when looking for associations between longitudinal profiles, the optimal solution could be to analyze absolute abundances. However, such data require spike-ins and are currently rarely available. Badri et al. (2018) have investigated normalization strategies and their effect in correlation analysis but for a single time point, while Metwally et al. (2018) proposed three normalization strategies that ignore the compositionality data problem. No method for longitudinal compositional data analysis has been proposed as yet. The proportionality measure proposed by Lovell et al. (2015) is a promising solution to reduce spurious correlations. However, it has not been developed for longitudinal problems, and the metric is not suitable in our context to perform variable selection. Instead, we chose to use the proportionality distance as a *post hoc* evaluation in our framework, not only to reduce potential spurious associations between profiles assigned in each cluster, but also to improve and help interpretation with respect to proportional and relative abundance of the profiles.

Finally, our framework does not include time delay analysis, even though dynamic delays between different types of molecules (e.g., DNA, RNA, or metabolites) can be expected. For example, 16S data describes the abundance of the microorganisms, with metabolites as the consequence of their activity, and performance as the macroscopic resulting output. Potential delays between these molecules can be detected using other techniques, such as the fast Fourier transform approach from Straube et al. (2017), and will be further investigated in our future work.

To summarize, we have proposed one of the first computational frameworks to integrate longitudinal microbiome data with other omics data or other variables generated on the same biological samples or material. The identification of highly associated key omics features can help generate novel hypotheses to better understand the dynamics of biological and biosystem interactions. Thus, our data-driven approach will open new avenues for the exploration and analyses of multi-omics studies.

DATA AVAILABILITY STATEMENT

Infant gut microbiota phylochip raw data can be found in Palmer et al. (2007). The microbiome and performance datasets for the bioreactor study can be found in Poirier and Chapleur (2018); metabolomic data are available on request. In-house scripts and code to conduct both case study analysis are available in a Github public repository: https://github.com/abodein/timeOmics

AUTHOR CONTRIBUTIONS

All authors contributed to the design of the study. AB and OC performed the statistical analyses. AB, OC, and K-ALC wrote the manuscript. All authors read and approved the submitted version.

FUNDING

Waste degradation study was supported in part by the Digestomic project funded by the French National Research Agency (ANR-16-CE05-0014). K-ALC was supported in part by the National Health and Medical Research Council (NHMRC) Career Development fellowship (GNT1159458). K-ALC and OC scientific travels were supported in part by the France-Australia Science Innovation Collaboration (FASIC) Program Early Career Fellowships from the Australian Academy of Science. AB and AD are supported by Research and Innovation chair L'Oreal in Digital Biology.

ACKNOWLEDGMENTS

We thank Angéline Guenne for analytical support with GC-MS analysis, Kodjovi Dodji Mlaga for the biological interpretations of the infant study, and Zoe Welham for proof-reading the manuscript. We 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.2019.00963/full#supplementary-material

REFERENCES

Äijö, T., Müller, C. L., and Bonneau, R. (2017). Temporal probabilistic modeling of bacterial compositions derived from 16s rrna sequencing. *Bioinformatics* 34, 372–380. doi: 10.1093/bioinformatics/btx549

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B (Methodol.)* 44 (2), 139–160. doi: 10.1111/j.2517-6161.1982.tb01195.x

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with graphlan. *PeerJ* 3, e1029. doi: 10.7717/peerj.1029

Badri, M., Kurtz, Z., Muller, C., and Bonneau, R. (2018). Normalization methods for microbial abundance data strongly affect correlation estimates. *bioRxiv*, 406264. doi: 10.1101/406264

Baksi, K. D., Kuntal, B. K., and Mande, S. S. (2018). 'time': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.* 9, 36. doi: 10.3389/fmicb.2018.00036

Bing, M., Forney, L., and Ravel, J. (2012). The vaginal microbiome: rethinking health and diseases. *Annu. Rev. Microbiol.* 66, 371–389. doi: 10.1146/annurev-micro-092611-150157

Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin biosynthesis. *Annu. Rev. Plant Biol.* 54, 519–546. doi: 10.1146/annurev.arplant.54.031902.134938

Breusch, T. S., and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econ.: J. Econom. Soc.* 47 (5), 1287–1294. doi: 10.2307/1911963

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, 121. doi: 10.1186/s13059-016-0980-6

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. doi: 10.1038/nmeth.f.303

Chapleur, O., Bize, A., Serain, T., Mazéas, L., and Bouchez, T. (2014). Co-inoculating ruminal content neither provides active hydrolytic microbes nor improves methanization of 13c-cellulose in batch digesters. *FEMS Microbiol. Ecol.* 87, 616–629. doi: 10.1111/1574-6941.12249

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics* 53, 406–413. doi: 10.1198/TECH.2011.08118

Dudek-Wicher, R. K., Junka, A., and Bartoszewicz, M. (2018). The influence of antibiotics and dietary components on gut microbiota. *Przeglad Gastroenterol.* 13, 85. doi: 10.5114/pg.2018.76005

Durbán, M., Harezlak, J., Wand, M., and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Stat. Med.* 24, 1153–1167. doi: 10.1002/sim.1991

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., et al. (2017). Frogs: find, rapidly, otus with galaxy solution. *Bioinformatics* 34, 1287–1294. doi: 10.1093/bioinformatics/btx791

Faust, K., Lahti, L., Gonze, D., DeVos, 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

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. doi: 10.1186/2049-2618-2-15

Fukuyama, J., Rumker, L., Sankaran, K., Jeganathan, P., Dethlefsen, L., Relman, D. A., et al. (2017). Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.* 13, e1005706. doi: 10.1371/journal.pcbi.1005706

Gavin, P., Mullaney, J., Loo, D., Lê Cao, K. A., Gottlieb, P., Hill, M., et al. (2018). Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for type 1 diabetes. *Diabetes Care* 41, 2178–2186. doi: 10.2337/dc18-0777

Gerber, G. K., Onderdonk, A. B., and Bry, L. (2012). Inferring dynamic signatures of microbes in complex host ecosystems. *PLoS Comput. Biol.* 8, e1002624. doi: 10.1371/journal.pcbi.1002624

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

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol.* 9, 244. doi: 10.1038/nrmicro2537

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465. doi: 10.1038/nature16942

Hoyos-Hernandez, C., Hoffmann, M., Guenne, A., and Mazeas, L. (2014). Elucidation of the thermophilic phenol biodegradation pathway *via* benzoate during the anaerobic digestion of municipal solid waste. *Chemosphere* 97, 115–119. doi: 10.1016/j.chemosphere.2013.10.045

Huang, D. S., and Zheng, C. H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

Hyndman, R. J., and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Comput. Stat. Data Anal.* 51, 4942–4956. doi: 10.1016/j.csda.2006.07.028

Jolliffe, I. (2011). *Principal component analysis*. Berlin Heidelberg: Springer. doi: 10.1002/0470013192.bsa501

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30, 513. doi: 10.1038/nbt.2235

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12, 118–123. doi: 10.1111/j.1462-2920.2009.02051.x

Lê Cao, K., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Stat. App. Genet. Mol. Biol.* 7 (1), 1–29. doi: 10.2202/1544-6115.1390

Lê Cao, K. A., Costello, M. E., Chua, X. Y., Brazeilles, R., and Rondeau, P. (2016a). Mixmc: Multivariate insights into microbial communities. *PLoS One* 11, e0160169. doi: 10.1371/journal.pone.0160169

Lê Cao, K. A., Costello, M. E., Lakis, V. A., Bartolo, F., Chua, X. Y., Brazeilles, R., et al. (2016b). Mixmc: a multivariate statistical framework to gain insight into microbial communities. *PloS One* 11, e0160169. doi: 10.1371/journal.pone.0160169

Limam, I., Guenne, A., Driss, M. R., and Mazéas, L. (2010). Simultaneous determination of phenol, methylphenols, chlorophenols and bisphenol-a by headspace solid-phase microextraction-gas chromatography-mass spectrometry in water samples and industrial effluents. *Int. J. Environ. Anal. Chem.* 90, 230–244. doi: 10.1080/03067310903267307

Liu, P., Qiu, Q., and Lu, Y. (2011). Syntrophomonadaceae-affiliated species as active butyrate-utilizing syntrophs in paddy field soil. *Appl. Environ. Microbiol.* 77, 3884–3887. doi: 10.1128/AEM.00190-11

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* 11, e1004075. doi: 10.1371/journal.pcbi.1004075

Luo, D., Ziebell, S., and An, L. (2017). An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics* 33, 1286–1292. doi: 10.1093/bioinformatics/btw828

Madigou, C., Lê Cao, K. A., Bureau, C., Mazéas, L., Déjean, S., and Chapleur, O. (2019). Ecological consequences of abrupt temperature changes in anaerobic digesters. *Chem. Eng. J.* 361, 266–277. doi: 10.1016/j.cej.2018.12.003

Metwally, A. A., Yang, J., Ascoli, C., Dai, Y., Finn, P. W., and Perkins, D. L. (2018). Metalonda: a flexible r package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome* 6, 32. doi: 10.1186/s40168-018-0402-y

Morris, A., Paulson, J. N., Talukder, H., Tipton, L., Kling, H., Cui, L., et al. (2016). Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term shiv infection. *Microbiome* 4, 38. doi: 10.1186/s40168-016-0183-0

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5, e177. doi: 10.1371/journal.pbio.0050177

Paulson, J. N., Talukder, H., and Bravo, H. C. (2017). Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. *BioRxiv*, 099457. doi: 10.1101/099457

Poirier, S., and Chapleur, O. (2018). Inhibition of anaerobic digestion by phenol and ammonia: Effect on degradation performances and microbial dynamics. *Data Brief* 19, 2235–2239. doi: 10.1016/j.dib.2018.06.119

Poirier, S., Desmond-Le Quéméner, E., Madigou, C., Bouchez, T., and Chapleur, O. (2016). Anaerobic digestion of biowaste under extreme ammonia concentration: identification of key microbial phylotypes. *Bioresour. Technol.* 207, 92–101. doi: 10.1016/j.biortech.2016.01.124

Quinn, T. P., Richardson, M. F., Lovell, D., and Crowley, T. M. (2017). propr: an r-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* 7, 16252. doi: 10.1038/s41598-017-16520-0

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66, 846–850. doi: 10.1080/01621459.1971.10482356

Ribicic, D., McFarlin, K. M., Netzer, R., Brakstad, O. G., Winkler, A., Throne-Holst, M., et al. (2018). Oil type and temperature dependent biodegradation dynamics-combining chemical and microbial community data through multivariate analysis. *BMC Microbiol.* 18, 83. doi: 10.1186/s12866-018-1221-9

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. doi: 10.1038/ismej.2017.107

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017). Mixomics: an r package for 'omics feature selection and multiple data integration. *PLoS Computat. Biol.* 13. doi: 10.1371/journal.pcbi.1005752

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

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Stat.* 11, 735–757. doi: 10.1198/106186002853

Rutayisire, E., Huang, K., Liu, Y., and Tao, F. (2016). The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. *BMC Gastroenterol.* 16, 86. doi: 10.1186/s12876-016-0498-0

Shields-Cutler, R. R., Al-Ghalith, G. A., Yassour, M., and Knights, D. (2018). Splinectomer enables group comparisons in longitudinal microbiome studies. *Front. Microbiol.* 9, 785. doi: 10.3389/fmicb.2018.00785

Shin, H., Pei, Z., Martinez, K. A., Rivera-Vinas, J. I., Mendez, K., Cavallin, H., et al. (2015). The first microbial environment of infants born by c-section: the operating room microbes. *Microbiome* 3, 59. doi: 10.1186/s40168-015-0126-1

Silverman, B. W., et al. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Stat.* 24, 1–24. doi: 10.1214/aos/1033066196

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). Diablo: an integrative approach for identifying key molecular drivers from multi-omic assays. *Bioinformatics* 35 (17), 3055–3062. doi: 10.1093/bioinformatics/bty1054

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787. doi: 10.1021/ac051437y

Straube, J., Gorse P, A. D., Huang, B., and Lê Cao, K. A. (2015). A linear mixed model spline framework for analysing time course omics data. *PLoS One* 10 (8), e0134540. doi: 10.1371/journal.pone.0134540

Straube, J., Huang, B. E., and Lê Cao, K. A. (2017). Dynomics to identify delays and co-expression patterns across time course experiments. *Sci. Rep.* 7, 40131. doi: 10.1038/srep40131

Straube, J., Lê Cao, K. A., and Huang, E., (2016). *lmms: Linear Mixed Effect Model Splines for Modelling and Analysis of Time Course Data*. R package version 1.3.3.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K. A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15, 569–583. doi: 10.1093/biostatistics/kxu001

Tenenhaus, A., and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika* 76, 257–284. doi: 10.1007/s11336-011-9206-8

Thursby, E., and Juge, N. (2017). Introduction to the human gut microbiota. *Biochem. J.* 474, 1823–1836. doi: 10.1042/BCJ20160510

Tibshirani, R. (1996). Regression shrinkage and selection *via the* lasso. *J. Royal Stat. Soc. Ser. B (Methodol.)* 58 (1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Torres, B., Porras, G., García, J. L., and Díaz, E. (2003). Regulation of the mhp cluster responsible for 3-(3-hydroxyphenyl) propionic acid degradation in escherichia coli. *J. Biol. Chem.* 278 (30), 27575–27585. doi: 10.1074/jbc.M303245200

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *J. Royal Stat. Soc.* 48, 269–311. doi: 10.1111/1467-9876.00154

Wang, K., Wang, B., and Peng, L. (2009). Cvap: validation for cluster analyses. *Data Sci. J.* 8, 88–93. 0904220071–0904220071. doi: 10.2481/dsj.007-020

Watkins, P. A., Moser, A. B., Toomer, C. B., Steinberg, S. J., Moser, H. W., Karaman, M. W., et al. (2010). Identification of differences in human and great ape phytanic acid metabolism that could influence gene expression profiles and physiological functions. *BMC Physiol.* 10, 19. doi: 10.1186/1472-6793-10-19

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534. doi: 10.1093/biostatistics/kxp008

Wold, H. (1975). "Path models with latent variables: The NIPALS approach." *Quantitative Sociology.* (New-York, USA.: Academic Press) 307–357. doi: 10.1016/B978-0-12-103950-9.50017-4

Yao, F., Müller, H. G., Wang, J. L., et al. (2005). Functional linear regression analysis for longitudinal data. *Ann. Stat.* 33, 2873–2903. doi: 10.1214/009053605000000660

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* 95, 601–619 doi: 10.1093/biomet/asn035.

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.

The handling editor and reviewer LA declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Genetics

REVIEW
published: 08 November 2019
doi: 10.3389/fgene.2019.00995

Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities

*Duo Jiang[1], Courtney R. Armour[2], Chenxiao Hu[1], Meng Mei[1], Chuan Tian[1], Thomas J. Sharpton[1,2] and Yuan Jiang[1]**

[1] Department of Statistics, Oregon State University, Corvallis, OR, United States, [2] Department of Microbiology, Oregon State University, Corvallis, OR, United States

The advent of large-scale microbiome studies affords newfound analytical opportunities to understand how these communities of microbes operate and relate to their environment. However, the analytical methodology needed to model microbiome data and integrate them with other data constructs remains nascent. This emergent analytical toolset frequently ports over techniques developed in other multi-omics investigations, especially the growing array of statistical and computational techniques for integrating and representing data through networks. While network analysis has emerged as a powerful approach to modeling microbiome data, oftentimes by integrating these data with other types of omics data to discern their functional linkages, it is not always evident if the statistical details of the approach being applied are consistent with the assumptions of microbiome data or how they impact data interpretation. In this review, we overview some of the most important network methods for integrative analysis, with an emphasis on methods that have been applied or have great potential to be applied to the analysis of multi-omics integration of microbiome data. We compare advantages and disadvantages of various statistical tools, assess their applicability to microbiome data, and discuss their biological interpretability. We also highlight on-going statistical challenges and opportunities for integrative network analysis of microbiome data.

Keywords: compositionality, heterogeneity, microbiome networks, multi-omics data integration, network analysis, normalization, sparsity

OPEN ACCESS

Edited by:
Himel Mallick, Merck, United States

Reviewed by:
Angela Re, Italian Institute of Technology, Italy
Yuehua Cui, Michigan State University, United States

****Correspondence:***
Yuan Jiang
yuan.jiang@stat.oregonstate.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 February 2019*
Accepted: *18 September 2019*
Published: *08 November 2019*

Citation:
Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ and Jiang Y (2019) Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. Front. Genet. 10:995. doi: 10.3389/fgene.2019.00995

INTRODUCTION

The microbiological sciences have undergone a research transformation in recent years as extensive volumes of microbiome data have been generated. By coupling environmental DNA sequencing procedures with bioinformatic and data analytic approaches, scientists have begun to disentangle the composition, diversity, and function of microbiomes (The Human Microbiome Project Consortium, 2012; Sunagawa et al., 2015; Thompson et al., 2017). However, the complexity of microbial systems, which frequently include diverse taxa and ecological covariates, continues to challenge the discovery of biological signal in these massive data sets. One common goal is to resolve how the microbiome influences or responds to its environment (Alivisatos et al., 2015; Blaser et al., 2016). To disentangle these mechanisms among the complex milieu of microbiome features, researchers have developed a rich array of analytical procedures, with one of the most widely used being microbiome network reconstruction.

Networks can be used to itemize interactions between community members, between communities, and between community members and some set of covariates (Follows et al., 2007; Faust et al., 2012; Gaulke et al., 2016; Tapio et al., 2017; Gould et al., 2018; Mandakovic et al., 2018). As a result, they offer a mapping of how information flows among the members of the microbiome or its environment (Röttjers and Faust, 2018). These networks have been most widely applied to microbiome taxonomic data and are traditionally assembled by correlating microbiome features and establishing linkages between features based on the significance or magnitudes of these correlations (Faust and Raes, 2012; Röttjers and Faust, 2018). Networks can then be visualized or analyzed using a variety of techniques to resolve, for example, taxa that potentially co-depend on one another, taxa that potentially compete with one another, or keystone taxa (Faust and Raes, 2012; Layeghifard et al., 2017). More analytically rigorous methods for inferring these taxonomic interactions have recently been developed to resolve the biologically relevant interactions and to account for unique statistical features of microbiome data (Dohlman and Shen, 2019).

While the analysis of networks representing microbe-microbe interactions has transformed our knowledge of how uncultured microbes potentially interact with one another in their environment, a small but growing number of studies increasingly leverage multi-omics networks to infer how microbial taxa interact with features of their environment (Kint et al., 2010; McHardy et al., 2013; Theriot et al., 2014; Morgun et al., 2015; Heintz-Buschart et al., 2016; Pfalzer et al., 2016; Maier et al., 2017). Microbiome multi-omics data involve collecting multiple types of high-dimensional biological data—including 16S, metagenomic, metatranscriptomic, metabolomics, etc.—from a microbiome sample and its environment or host. While these approaches often remain relatively expensive, technological transformations continue to reduce the cost of generating diverse data constructs, which, in turn, increases the rate at which researchers can apply these multi-omics approaches. This increased accessibility is fortunate, as the integration of multi-omics data holds potential to resolve functional mechanisms of the microbiome (Rodrigues et al., 2018; Wang et al., 2019). For example, these data integrative networks can clarify how changes in the relative abundance of a taxon relates to the expression of genes across a microbial community (i.e., the metatranscriptome), the pool of metabolites, or the phenotype of the microbiome's host. However, there remain relatively few tools that investigators can rely on to integrate and understand these data.

Multi-omics network integration offers an opportunity to resolve how specific members of the microbiome functionally relate to specific environmental features, which, in turn, helps researchers key in on pathways of information flow that may ultimately transform our ability to manipulate, rescue, or mimic microbiomes. However, their application remains nascent. Most studies in this area thus far apply measures of correlation (such as Spearman's rank correlation) to resolve microbial taxa that correlate with specific environmental or host features. This approach has specifically been used to clarify how gut microbial abundance relates to the pool of intestinal metabolites (McHardy et al., 2013), discern possible connections between mucosal bacterial abundance and intestinal gene expression in association with inflammatory bowel disease (Morgan et al., 2015), resolve which specific microbes on the human skin may produce metabolites of interest (Bouslimani et al., 2015), and uncover how ocean microbes express transcripts (Aylward et al., 2015). However, this relatively simplistic statistical approach does not necessarily meet the assumptions of microbiome data or address the needs of the problems that arise from such data and may yield inappropriate conclusions.

To promote the innovation of statistical approaches that are more appropriate and specific for microbiome multi-omics network analysis, we present a comprehensive review of the currently available network-based statistical methods and discuss their application to multi-omics data integration. In addition, we consider the unique features of microbiome data and microbiome multi-omics data integration and further explore the reviewed network-based statistical methods in terms of their appropriateness and limitation when applied to microbiome multi-omics data integration. At the end, we conclude with remarks on the major challenges and research opportunities in the innovation of statistical approaches for microbiome multi-omics network analysis.

OVERVIEW OF NETWORKS

Network data structures are often complex and involve rich and unusual terminology. In this section, we orient readers to basic concepts and terms associated with network data science, with the goal of improving comprehension of the subsequent discussion of network-based statistical approaches (Section "Review of Available Network-Based Procedures").

Networks, which are also called graphs, are useful data structures for examining how components of a system interact with or relate to one another. These interactions are commonly derived using statistical approaches that reveal associations between pairs of components and are further illustrated graphically as edges that connect pairs of nodes that represent the components of a system. Networks can also represent empirical interactions between components that have been experimentally validated. However, in the case of microbiome research, limitations in the number of cultured taxa and the complexity of most microbial communities restrict the application of such empirical approaches. Networks have been effectively used in a variety of fields. Examples include infectious disease research (Silk et al., 2017), social interaction analysis applied to marketing (Liu et al., 2019) and political science (Cranmer et al., 2017), analysis of neuroimaging data (Fujita et al., 2017), information flow through the internet (Dorogovtsev and Mendes, 2003), genomics data analysis (Kleaveland et al., 2018). In microbiome science, network data structures have been used in a variety of contexts (as reviewed by Faust and Raes, 2012, and Layeghifard et al., 2017), including efforts to evaluate interactions between members of a microbial community (Faust et al., 2012), associate taxa with metabolite production (Bouslimani et al., 2015), and determine which taxa interact with host bile acid metabolism (Theriot et al., 2016).

Networks adopt a variety of terms and properties, some of which we define here to orient readers. The components of the system being modeled by a network are represented as nodes or vertices. In microbiome research nodes can be biological features such as microbial taxa, genes, metabolites, and proteins. Nodes may also represent environmental or host features, such as pH and markers of immune status. The presence of an edge between a pair of nodes indicates an association between the nodes, such as a correlation between the abundance of two taxa. Such edges may suggest a dependency between the taxa by indicating, for example, that when one taxon increases in abundance, the other taxa do as well possibly due to cross-feeding. We note that an inferred edge itself does not imply a causal dependency between the features, the inference of which requires a controlled experiment. If the associations differ in strength, edges can be weighted to illustrate the strength of association and guide interpretation. The distinction between positive and negative associations can also be captured by weights of different signs. In some cases, the interactions being modeled by a network are directed, meaning that they indicate that the change to one component causes a change in another connected component. In such instances, such directed network edges are represented by arrows and can be used to depict the cause and effect relationships among components. It is worth noting that causality can be challenging or impossible to infer in many genomic investigations depending on the study design. In those cases, the directionality of the relationship might be pre-specified based on knowledge to construct a bipartite network (e.g., in some regression models, see Section "Regression-Based Methods") or inferred using the data in a probabilistic framework as a way of representing the information propagation in the system (e.g., in Bayesian networks, see Section "Bayesian Networks").

In this article, our main interest is in the problem of estimating or constructing a network by integrating two or more types of omics data including microbiome data. In the rest of this article, the variables representing the components corresponding to each data type will be referred to as "features." Variables from different types of omics data will be said to belong to different "feature types." Examples of feature types include but are not limited to microbiome taxonomic, transcriptomic, and metabolomic features. The corresponding features within these feature types may include the abundance of a microbial taxon, the expression level of a gene, and the concentration of a metabolite. Depending on the scientific question of interest and the analytical approach used, there are various types of networks that can be constructed based on multi-omics data. When considering associations between distinct feature types, a bipartite network can be used where the edges are drawn between nodes of different types (as reviewed by Pavlopoulos et al., 2018). Alternatively, it is possible to construct a network among features of a single type where data from another type are incorporated in the analysis as additional information or covariates to improve the estimation of the network. Examples of this approach include studies conducted by Li et al. (2012) and Chun et al. (2013), a more detailed discussion of which can be found in Section "Methods Based on Graphical Models".

Once networks are estimated from the data, there are numerous metrics that can be quantified on the networks to summarize the overall structure of the system. One of the primary metrics used is degree, which is the count of edges that connect one node to all the others. Nodes with higher degree represent features that are relatively highly connected to other features in the system being modeled. Such nodes may have more influence on the system's dynamics and may represent, for example, keystone taxa in a community. Most real-world networks have a right-skewed degree distribution where most vertices have low degree, and few have high degree. When the degree distribution monotonically decreases over its entire range, it has a power-law distribution and is referred to as a scale-free network. In a scale-free network, some nodes can have significantly higher degree than others. Such nodes are often referred to as "hubs" because they are strong participants of the interactions in the network. Another way to identify important nodes is through measures of betweenness. To calculate betweenness, the shortest path between each pair of nodes in the network is first identified. Then, the betweenness for each node is measured as the number of times the node in question lies in the shortest path between two other nodes. Nodes with high betweenness are potentially influential in the network since they come between many pairs of nodes. Nodes with high betweenness can also have high degree; however, that is not always the case. High-betweenness nodes are often interpreted as bottlenecks of the information flow in the network. Various other topological properties of the network can also be assessed to glean interesting biological insights into a system, such as modularity, which aims to identify clusters of nodes densely connected to each other, with relatively low connectivity to the rest of the network. We refer the readers to the papers of Newman (2010), Ma'ayan (2011), and Charitou et al. (2016) for more in-depth discussions of topological analysis techniques. In this review, we will focus on the statistical estimation of networks instead of the topological analysis of an estimated network.

REVIEW OF AVAILABLE NETWORK-BASED PROCEDURES

In recent years, integrative network analysis has increased in popularity, particularly for multi-omics data sets. The statistical methods utilized in these analyses lend perspective to how microbiome multi-omics networks can be inferred. In this section, we review network-based statistical methods with an emphasis on their applications to multi-omics data integration. We categorize commonly adopted methods into six types and present a detailed review of each type. **Table 1** provides a summary and a comparison of the six types of methods alongside software packages that enable their implementation.

Marginal Correlation Analysis

The most commonly applied statistical method for constructing biological networks is marginal correlation analysis. In this analysis, the relationship between two biological features, such as genes, transcripts, proteins, metabolites, and microbes, is described by the correlation of their expression, concentration, or abundance levels inferred from multiple statistically independent observations, such as biological replicates or samples. Technically,

TABLE 1 | Summary of available network-based procedures.

Method type	Network type	Representative methods (software: packages)	Advantages	Disadvantages
Marginal correlation analysis	Undirected	Pearson's correlation, Spearman's rank correlation, Kendall's tau (R: base); Local similarity analysis (Linux: ELSA); WGCNA (R: WGCNA)	Easy to implement; nonparametric options available.	Subject to spurious findings due to confounding.
Dimension reduction methods	Typically undirected	PCA (R: base); CCA (R: CCA); PLS (R: pls); CIA (R: ade4); Sparse CCA, Sparse multiple CCA (R: PMA); Sparse PLS (R: spls); Sparse CIA (R: pCIA); Kernel PCA, kernel CCA (R: kernlab)	Can be used to construct networks linking modules of features.	Poor interpretability because each node represents multiple, if not all, features.
Regression-based methods	Directed or undirected	Linear and generalized linear models (R: base); Linear and generalized linear mixed models (R: nlme, lme4); Regularized regression: Lasso, ridge, elastic net (R: glmnet), SCAD, MCP (R: ncvreg), Group lasso, group elastic net, group SCAD, group MCP (R: grpreg); Regularized multivariate regression: Graph-guided fused lasso (R: GFLASSO), remMap (R: remMap), Reduced-rank regression (R: rrpack)	Easy to incorporate covariates; a large number of statistical methods and software tools are available.	Need to specify each feature as either a response variable or a predictor.
Graphical models	Undirected	Graphical lasso (R: glasso, huge); Neighbourhood selection (R: huge); Joint graphical lasso (R: JGL); Conditional graphical models Covariated-adjusted graphical models (R code: caPC)	Conditional dependency captures direct biological interactions more effectively than methods based on marginal correlations.	Most methods assume a multivariate normal distribution.
Bayesian networks	Directed	CONEXIC (Linux: CONEXIC); QTLnet (R: qtlnet); Bayesian Network Prior (MATLAB: BNP); Search-and-score approaches, constrain-based approaches (R: bnlearn)	Links more directly related to causality; ability to incorporate prior knowledge; possibility to handle data following disparate distribution types.	Current methods do not scale well to massive data sets.
Network integration	Undirected	GeneMania (Cytoscape/Web: GeneMANIA); SNF (R: SNPtools); DCA (MATLAB: Mathup)	Often simple to implement; ability to borrow information from multiple networks.	Individual networks that serve as the input of the methods must be reliably estimated; a shared biological mechanism is assumed.

this relationship can be quantified by any statistical measure of correlation, including but not limited to Pearson's correlation, Spearman's rank correlation, and Kendall's tau, as long as the approach is meaningful for a given biological context. Marginal correlation analysis is also useful when integrating multiple biological feature types (e.g., genes, transcripts, and proteins) to uncover relationships across feature types (Heintz-Buschart et al., 2016; Bakker et al., 2018; Frost and Amos 2018; McGrail et al., 2018).

Marginal correlation analysis can also be extended to observations that are statistically dependent. For example, consider the case wherein two biological features are observed over time (i.e., two time series of measures). One might want to assess the correlation of the features across the time series. In this case, it is essential that the correlation measures account for the longitudinal nature of the observations. One approach to this problem is the so-called local similarity analysis of two time series (Ruan et al., 2006). In this approach, both time series are first transformed separately to their normal scores. Then, for any subsequence of the first time series starting from the beginning, all subsequences of the same length from the second time series are identified within some predefined time delay. Pearson's correlations are then calculated between each pair of subsequences across the two time series. Finally, the local similarity score is defined as the maximum correlation for all such possible pairs of subsequences, aiming to find associations with possible delays between the two time series. Local similarity analysis has proven useful for detecting co-varying pairs of microbes as well as the association between a microbe and an environmental factor (e.g., temperature), especially when the variations between features are not synchronous (Ruan et al., 2006).

While the abovementioned methods are purely data-driven, other methods construct biological networks based on both statistical correlations and existing biological knowledge. For example, to create a bipartite network describing the relationship between mRNAs and miRNAs, Gade et al. (2011) combined two p-values for each pair of mRNA and miRNA expression values: (a) a p-value measuring the statistical correlation of the observed data and (b) a p-value obtained from an existing database of miRNA-target predictions (e.g., miRBase) (Griffiths-Jones et al., 2008). The authors applied a truncated product method of combining p-values (Zaykin et al., 2002), which they then transformed to weights and viewed as the adjacency matrix of a bipartite network describing the relationship between mRNAs and miRNAs.

In order to produce a biological network that facilitates meaningful interpretations, studies often only include correlations in the network that manifest correlation coefficients whose absolute value exceeds a threshold, which is usually arbitrarily determined, or if its associated p-value is less than a significance

level such as 0.05. In the latter case, some applications simply use the raw p-values, which tend to yield excessive false positive edges, while other applications more carefully control false positives by adjusting the p-values with a multiple testing correction for familywise error rate (FWER) or false discovery rate (FDR). A biological network is then constructed by connecting those pairs of biological features with a statistically robust correlation and leaving all other pairs unconnected.

The abovementioned thresholding procedure produces a biological network that is unweighted, in the sense that an edge either exists or not between any pair of nodes. Weighted networks based on marginal correlation analysis have also attracted recent attention, such as in the case of Weighted Gene Co-expression Network Analysis (WGCNA) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008). In this method, an edge in a network is weighted by a soft thresholding function of the inferred correlation (e.g., the sigmoid function, the power adjacency function, etc.) on a continuous scale. Many topological analysis methods have also been extended from unweighted networks to weighted networks, such as node connectivity (Barrat et al., 2003; Amano et al., 2018), network modules (Newman, 2004; Li et al., 2011; Lecca and Re, 2015), clustering coefficient (Opsahl and Panzarasa, 2009), and scale-free topology (Tan and Lei, 2013; Zhang et al., 2015). Because weighted networks encode additional information in the form of connection strengths as compared to unweighted networks, weighted networks have been shown to be a useful option for many biological datasets, including but not limited to microarray data (Kadarmideen et al., 2011; Mohammadnejad et al., 2019), single cell RNA-Seq data (Xue et al., 2013), DNA methylation data (Horvath et al., 2012; Wang et al., 2016), and microbiome data (Tong et al., 2013; Li et al., 2019).

Marginal correlation analysis is probably the most commonly used method to infer biological networks due to its computational simplicity. However, the approach is limited by the fact that it can only infer relationships between pairs of biological features and does not consider how the observed relationship may depend upon other variables or features. As a result, marginal correlation analysis can lead to spurious correlations: two features that independently interact with a third, but not with one another, may appear to correlate. Therefore, marginal correlation analysis is known to be prone to false positives when seeking to identify direct interactions or causal effects among the features. It is important to keep this limitation in mind and to critically assess the risk of confounding factors before drawing conclusions about biological interactions that result from marginal correlation analysis.

Dimension Reduction Methods

Dimension reduction, such as the widely used method principal component analysis (PCA), is a useful statistical tool that aims to reduce the dimension of a set of variables while retaining as much information from the original data as possible. It is also useful when the relationships between two feature types are investigated, in which case data associated with each feature type are reduced to a lower dimension in a way that captures as much association between the two feature types as possible. We refer the readers to the review papers of Burges (2009) and Engel et al. (2011) as two statistical reviews on dimension reduction and to the review paper of Meng et al. (2016) as a review on the application of dimension reduction to the integrative analysis of multi-omics data.

Commonly used dimension reduction tools include canonical correlation analysis (CCA), partial least square regression (PLS), and co-inertia analysis (CIA) (Meng et al., 2016). These tools share the same goal of summarizing the variables in each feature type by using a small number of linear combinations so as to maximize the association between the two feature types as demonstrated by these linear combinations. Different measures of association correspond to different tools in this category. More specifically, CCA uses Pearson's correlation to capture the association between two linear combinations (or equivalently, all linear combinations are normalized to have a unit variance), PLS uses covariance to quantify the association with the constraint that the linear combination from one feature type has a unit variance, and CIA uses covariance to represent the similarity with no variance constraint. CCA, PLS, and CIA have all been applied to infer biological networks from multi-omics data. For example, CCA was used to construct gene co-expression networks by considering linear combinations of gene expression at the exon or base pair level for each gene obtained from an RNA-seq dataset (Hong et al., 2013). In this study, the authors then calculated the canonical correlation between each pair of genes, ranked the correlations based on their magnitude, and constructed a co-expression network by retaining a predetermined percentage of edges. In other studies, CIA was applied to mRNA and microRNA data to determine which microRNAs regulates gene expressions (Jovanović et al., 2014) as well as to microbiome and metabolomic data sets to understand the impact of a short-term increase in dietary fiber intake on the gut microbial community (Tap et al., 2015). PLS has also been utilized in multi-omics studies, for example, to analyze the associations between biomarkers for insulin sensitivity and a variety of omic data, including gut microbiota, adipose gene expression, and metabolomic data (Dao et al., 2019).

These methods suffer from a few limitations, which recent efforts have sought to overcome. The first limitation stems from the fact that a linear combination found by CCA, PLS, and CIA tends to include every variable under consideration, albeit with varying weights. This tendency to include every variable results in poor interpretability as it can be difficult to determine which variables contribute to the canonical correlations and which do not. Therefore, a desirable extension is to introduce sparsity to the linear combinations, where the coefficients for variables with less contribution are shrunk to zero. Recent methods that apply such a strategy include sparse canonical correlation analysis (SCCA) (Parkhomenko et al., 2007; Waaijenborg et al., 2008; Parkhomenko et al., 2009; Witten and Tibshirani, 2009; Witten et al., 2009; Hardoon and Shawe-Taylor, 2011; Suo et al., 2017), sparse partial least squares (SPLS) (Lê Cao et al., 2008; Chun and Keleş, 2010; Chung and Keles, 2010; Lee et al., 2011), and sparse co-inertia analysis (SCIA) (Min et al., 2018). These methods try to balance between maximizing the correlation

between linear combinations defined for different feature types and minimizing the number of variables included in each linear combination. These methods share the same basic idea of incorporating variable selection techniques, such as lasso and elastic net (Tibshirani, 1996; Zou and Hastie, 2005), into traditional dimension reduction methods. As a result, these methods produce a sparse linear combination for each group of variables, although they each differ in either the problem formulation or computational details. These methods have been used to integrate SNP and gene expression data with the goal of identifying a group of SNPs that explain the variation in gene expression across a group of genes while keeping the group sizes sufficiently small to aid biological interpretation (Parkhomenko et al., 2007; Parkhomenko et al., 2009).

Another limitation of the traditional dimension reduction tools is that they can only consider two feature types, i.e., two groups of variables. Extensions of SCCA have been proposed to accommodate the analysis of multiple groups of variables (Witten and Tibshirani, 2009; Tenenhaus et al., 2014). Meng et al. (2014) proposed the multiple CIA method and used it to integrate transcriptomic, proteomic, and metabolomic data. All of these methods aim to find a linear combination from each group of variables so as to maximize the sum of squared pairwise correlations or the sum of squared covariances between each linear combination and a synthetic axis that is also parametrically optimized.

The third limitation of the traditional dimension reduction tools is that they only replace the original features by their linear combinations. Nonlinear dimension reduction tools have also been proposed to overcome this limitation such as kernel-based dimension reduction methods including kernel principal component analysis (KPCA) (Schölkopf et al.,1997), kernel canonical correlation analysis (KCCA) (Lai and Fyfe, 2003), and kernel fusion methods (Daemen et al., 2009). For example, Reverter et al. (2012) applied KPCA to classify disease types using the kernel principal components estimated from gene expression profiles. Daemen et al. (2009) proposed a kernel fusing method for clinical decision support that transforms multi-omics data into a linear combination of their corresponding kernel matrices and implements a classifier based on the combined result.

A common feature of the aforementioned dimension reduction tools for multi-omics data integration is that they are all based on the integration of two or more types of observed data. They are thus sometimes referred to as data-driven methods. Another class of dimension reduction tools try to integrate the observed data with external knowledge and are therefore called knowledge-driven methods. As an example, Yang et al. (2009) proposed a method called knowledge-based matrix factorization (KMF). In this study, the authors used KMF to build a gene co-expression network based on pairwise correlations between gene expression levels while incorporating existing pathway information from external databases such as Gene Ontology (GO) (Gene Ontology Consortium 2004). To incorporate this external knowledge, KMF finds the best low-rank factorization of the correlation matrix so that it is decomposed into the product of three matrices. The left and right matrices are transpose of each other and they approximate the membership of genes in pathways, while the center matrix captures the relationship between the pathways. This procedure allows KMF to construct a gene-gene correlation network whose structure is consistent with external pathway information while also identifying interactions between the pathways.

In summary, dimension reduction methods look for a combination of the features to represent each feature type while maximizing the correlation or covariance between the resulting combinations. Therefore, dimension reduction methods can be regarded as a multivariate extension of marginal correlation analysis. As a result, these methods are subject to the same pitfall that marginal correlation analysis faces (see Section "Marginal Correlation Analysis"); for example, they may lead to spurious correlations caused by confounding factors. In addition, although sparse versions of dimension reduction methods have been developed, lack of interpretability remains a limitation because each combination includes multiple, if not all, biological features in a group, and thus, the inferred relationships cannot be attributed to a specific pair of features.

Regression-Based Methods

Network inference in multi-omics data have also been formulated as a regression problem. In this case, a series of regression models are fitted by taking one feature type as the response variable and another type as the predictor variable. Associations identified by these regression models are often interpreted as a directed relationship in which the feature type serving as the predictor is considered to affect or explain the feature type serving as the response. However, this inferred effect does not necessarily demonstrate a cause and effect relationship among the variables. For example, to assess the extent to which mRNA abundance was able to explain protein abundance, Nie et al. (2006b) fitted a linear model for each protein-mRNA pair with the former as the response and the latter as a predictor, incorporating multiple sequence features as additional covariates. For noncontinuous data, generalized linear models such as Poisson regression have also been employed to elucidate interactions between genomic features (Nie et al., 2006a). More recently, Yuan et al. (2018) proposed a regression model that aims to infer gene regulatory networks by incorporating DNA methylation and copy number variation as well as their interactions. Regression-based methods have also been used to integrate other types of multi-omics data. Recent examples include a somatic eQTL analysis using linear regression to model the association between gene expression and the mutation status of linked loci while accounting for various covariates including DNA methylation and gene copy number variation (Zhang et al., 2018). Moore and Hoen (2019) discussed the use of the regression framework to analyze RNA-protein interactions.

As opposed to considering a single predictor at a time, each regression model can also simultaneously include a large number of predictors, possibly from multiple feature types, to identify a set of variables that best predict the response. Typically, in these methods, a feature type of interest is regarded as the response data, with the other feature types regarded as the explanatory data. In each regression model, one feature is taken as the response

variable, which is fitted against all variables in the explanatory data as predictors. The resulting high dimensionality leads to an underdetermined regression problem and thereby renders ordinary least squares and maximum likelihood estimation ill-posed. Therefore, variable selection techniques are needed to estimate the model parameters.

Regularized regression, the most representative method being lasso (Tibshirani, 1996), is commonly used for variable selection to overcome these limitations (as reviewed by Bickel and Li, 2006, and by Wu et al., 2019, for its application to multi-omics integration). In this case, a penalty term is incorporated in the usual least squares or maximum likelihood objective function in order to shrink some of the set of parameter estimates to zero, hence inducing sparsity in the regression coefficients. This strategy achieves variable selection and parameter estimation simultaneously. Each coefficient estimated to be nonzero is then represented by an edge in the network between the associated predictor and the response. There have been many applications of this approach to multi-omics studies. For example, Kim et al. (2014) and Yuan et al. (2018) estimated networks between DNA methylation, copy number variation, and gene expression based on a set of regularized linear regressions where separate L1 penalties were imposed on the three feature types. Qin et al. (2014) integrated ChIP seq and transcriptome data to infer gene regulatory networks using a regularization method where the L1 penalty is replaced by L0 and L0.5 penalties.

Another type of regression-based method for integrative network inference uses a technique called multivariate regression (Kim et al., 2009; Peng et al., 2012), which includes a multivariate response (i.e., multiple response variables) in a single model. When a multivariate response is modeled against a set of predictors, the unknown coefficients come in the form of a matrix, where an entry is assigned to relate each response variable to each predictor. Constraints are often imposed either on the sparsity or the rank of this coefficient matrix, or both, to ensure that the model can be fitted despite the limited sample size in comparison to the number of parameters. Applications of this approach to multi-omics data usually combine variables from one feature type, which serves as the multivariate response, while another type of omics features serve as the predictor variables. Like methods based on univariate regression, a directed network can be constructed with edges corresponding to nonzero coefficients. However, unlike univariate methods which involve a large number of separate regression models, multivariate regression only fits one joint model, which allows more realistic modeling and simplified understanding of the biological mechanisms *via* sparsity and rank constraints. For example, Goh et al. (2017) proposed a multivariate regression method, which was used to fit time-course mRNA data for >500 genes against binding information of the target genes for >100 transcription factors. Sparsity and low-rank constraints were imposed to account for the fact that many transcription factors are not related to the genes and the samples are correlated due to the study design.

Regression-based methods are widely used to construct biological networks mainly because they are relatively straightforward to implement. Compared with marginal correlation analysis and dimension reduction methods, regression models have the advantage of being able to incorporate relevant covariate information. A regression framework is also equipped with many well-studied statistical tools to flexibly handle specific analytical needs. For example, random effects can be incorporated to account for inter-sample correlation between samples due to study design (Zhang et al., 2013) and to correct for data heterogeneity due to unobserved confounders (Furlotte et al., 2011). The regression-based approach is also empowered by the recent statistical developments in penalized regression to handle high-dimensional data. However, most regression-based methods entail that each feature (or feature type) is identified as either a response variable or a predictor, which can be a nontrivial choice to make especially when the underlying biology is poorly understood for the system being studied.

Methods Based on Graphical Models

Gaussian graphical models are widely applied in network analysis (as reviewed by Drton and Maathuis, 2017). Specifically, in a multivariate Gaussian distribution, two variables are statistically independent conditional on all the other variables if and only if the corresponding entry in the inverse covariance matrix of the distribution is zero. Then, to construct a network with each edge representing the conditional dependence between two features given all other features, it is equivalent to identify the nonzero entries of the inverse covariance matrix for the multivariate Gaussian distribution. In reality, the data are often high-dimensional with more variables than samples, which leads to a degenerate sample covariance matrix and makes the estimation of the inverse covariance matrix challenging.

There are two major statistical approaches for estimating the inverse covariance matrix in the high-dimensional Gaussian graphical model: the neighborhood selection method (Meinshausen and Bühlmann, 2006) and the graphical lasso method (Yuan and Lin, 2007; Friedman et al., 2008). Both methods yield a sparse estimator of the inverse covariance matrix, whose nonzero entries can be used to construct a network that denotes the conditional dependency between the variables in the Gaussian graphical model. To apply Gaussian graphical models to the integration of multi-omics data, a naive strategy combines all variables from multiple feature types into one vector, which is assumed to follow a multivariate normal distribution (Shin et al., 2014). However, this approach effectively treats all variables as exchangeable, and, in turn, ignores the potentially important information about their group structure.

One typical application of Gaussian graphical models to multi-omics data is the joint Gaussian graphical model, which simultaneously estimates multiple graphical models under some constraints among them. The constraints are often determined by some prior knowledge for the multiple inverse covariance matrices such as their similarity in magnitudes or sparsity or the membership of nodes in biological pathways (Guo et al., 2011; Danaher et al., 2014; Kim et al., 2017). This idea has been applied to find biological networks from different groups simultaneously, e.g., disease subtypes or

experimental conditions. For example, Kim et al. (2017) used a joint Gaussian graphical model to estimate multiple mRNA expression networks from different datasets. Zhang et al. (2016) further extended the idea of joint graphical models to a two-dimensional joint graphical lasso model. This model imposed a joint penalty function to simultaneously estimate two gene expression networks that are patient group-specific from gene expression profiles collected from different data generation platforms. After obtaining the gene networks, the differential networks between the two patient groups were constructed by calculating the differences of dependencies between two group-specific networks (i.e., one differential network for each platform).

Bayesian inference based on joint Gaussian graphical models has also been used to construct networks by applying a G-Wishart prior on the inverse covariance matrix (Peterson et al., 2015). In this particular case, a Markov random field prior was imposed to encourage common edges between joint graph structures. This procedure enabled the identification of which groups have a shared network structure by placing a spike-and-slab prior on parameters which measure network relatedness.

Conditional graphical models represent another class of graphical model approaches that are useful for solving data integration problems. Different from the traditional graphical models, the conditional graphical model incorporates an additional conditioning step to remove spurious dependence that may be caused by common external factors. For example, two genes may depend on each other only because they are regulated by the same DNA markers and have no relationship otherwise. Along this research direction, Li et al. (2012) proposed a method which infers such a conditional graphical model in two steps. It first estimates the conditional covariance matrix and then uses penalized maximum likelihood to obtain the inverse conditional covariance estimator. The authors used their method to define a gene expression network conditional upon eQTL data. Moreover, Chun et al. (2013) extended the same idea to multiple conditional graphical models, allowing the integration of gene expression data from different sources, say, heart and fat tissues. Other similar research includes the covariate-adjusted graphical models that use genetic markers (SNPs) as covariates to correct both false positives and false negatives in gene regulatory networks (Cai et al., 2013; Gao and Cui, 2015). In these methods, the effect of genetic variation is estimated in the first step. Then, the graphical structure is estimated in the second step while adjusting for the genetic effects.

Like graphical lasso, most joint or conditional graphical models incorporate the sparsity assumption to tackle the high dimensionality problem in the context of inverse covariance matrix estimation, but often rely on the assumption of a multivariate Gaussian distribution. Zhang et al. (2017) is one of the few studies that estimate the inverse covariance matrix under a mixed model that includes different biological feature types by accommodating both discrete and continuous variables. Due to the computational complexity of discrete variables, the authors used the pseudo-likelihood method instead of the usual likelihood method for parameter estimation. In spite of these innovations, methods based on graphical models still need to account for the unique characteristics of microbiome data when applied to microbiome multi-omics data integration (Section "Unique Challenges of Microbiome Multi-Omics Network Analysis").

Bayesian Networks

Like Gaussian graphical models, Bayesian networks are probabilistic graphical models and are increasingly used as a statistical and machine learning tool for analyzing genomic data. In a Bayesian network, a graph with directed edges is used to represent the conditional relationships in the joint probability distribution of a set of variables: for each variable X, given its parent variables (i.e., nodes pointing to X), X only affects its child variables (i.e., nodes pointed to by X) and is conditionally independent of all other variables. These conditional independence constraints serve to cut down, frequently substantially, the number of parameters needed to jointly model the variables. We refer the readers to a review paper (Koski and Noble, 2014) for a more thorough introduction to Bayesian networks.

In the past decade, Bayesian networks have seen many applications in genomic data integration. For example, Akavia et al. (2010) introduced an algorithm based on Bayesian networks (CONEXIC) to identify driver mutations in cancer by integrating gene expression data with matched copy number data. QTLnet (Chaibub-Neto et al., 2010) is a method that uses a Bayesian network that includes both phenotype and genotype variables as nodes to jointly estimate the causal network between multiple phenotypes and their respective genetic architecture. In order to improve the recovery of gene interaction networks based on experimental data, Isci et al. (2014) proposed a hierarchical method called BNP where a Bayesian network is nested within a classical Bayesian modeling framework. This approach enables the incorporation of rich external knowledge about gene interactions as the prior information in the Bayesian inference procedure. More recently, Khanna et al. (2018) applied Bayesian network to elucidate the interplay between genotype information, neuroimaging measurements, and clinical data to help uncover biological mechanisms underlying Alzheimer's disease.

The Bayesian network approach has several appealing advantages when applied to multi-omics data analysis. First, because of the structure of the underlying probabilistic model, Bayesian networks are usually considered akin to directed networks, in such that causal relationships are often inferred among nodes. In particular, network edges are often interpreted to represent how information propagates between variables or components in a biological process. We note that, although causal interpretation of Bayesian networks is appealing and widespread, there have been growing skepticism over the liberal use of such interpretation because a Bayesian network does not guarantee causality (Korb and Nicholson, 2008). Second, Bayesian networks can incorporate prior knowledge about plausible relationships among variables within or between feature types (Ni et al., 2014). Third, Bayesian networks may be set up in a way that allows for simultaneous modeling of variables following different types of distributions. For example, Chaibub-Neto et al. (2010) modeled a Bayesian network where the nodes consist of a mixture of

continuous phenotype variables and discrete genetic variables. The ability to handle disparate data types is an attractive feature as multi-omics studies frequently involve feature types that are more appropriately modeled using different distributions such as continuous, count, and binary data.

However, a major challenge limiting the use of Bayesian networks in genomic studies is its steep computational cost. The estimation of the structure of a Bayesian network usually involves the optimization of a complicated objective function over a large, nonconvex search space. As the number of variables increases, the computational burden increases super exponentially. Consequently, in most applications of Bayesian networks to multi-omics data, either only a small to moderate number of omics variables are considered or dimension reduction techniques are applied to reduce the number of variables before implementing Bayesian networks.

Network Integration

A key goal of multi-omics data integration is to create a comprehensive view of a biological process from diverse types of omics data. Network integration approaches seek to solve this problem by integrating multiple, distinct biological networks assembled from different data types. There are many network integration strategies and we review below a representative subset of these approaches.

One approach to this problem, as illustrated by the method GeneMANIA (Mostafavi et al., 2008), is to build a composite association network by taking a weighted average of multiple association networks between features, such as genes, where the weights are selected based upon the composite network's ability to reconstruct referential characteristics of the features. For example, GeneMANIA uses ridge regression (Hoerl and Kennard, 1970) to find the weights of individual association networks to minimize the difference between the composite network and a target network constructed from known gene functions (such as GO functional categories), while incorporating the prior information of the weights in the ridge penalty.

Diffusion component analysis (DCA) (Cho et al., 2015; Wang et al., 2015; Cho et al., 2016) is another network integration method that targets heterogeneous networks with different connectivity patterns. In DCA, the diffusion state of each node is analyzed with the random walk with restart (RWR) method and is stored as a probability simplex that represents the probabilities that an RWR that starts at one node will end up at another node in equilibrium. A similar diffusion state between two nodes implies that the nodes are in similar positions within the network with respect to other nodes. Next, the node-specific diffusion state in individual networks are represented by two low-dimensional latent vectors: one that is shared across all networks and another that encodes the intrinsic topological property using multinomial logistic models. These shared low-dimensional node-specific latent vectors represent the homogeneous topological property across the network and can be used in other machine learning methods to derive further insights of the nodes. DCA has been applied to the functional analysis of genes (Cho et al., 2016) and drug-target interaction network (Luo et al., 2017).

While GeneMANIA and DCA integrates networks of features, similarity network fusion (SNF) (Wang et al., 2014) constructs a merged network between objects (e.g., biological samples) by combining multiple features types measured for each object. In particular, SNF first creates a network for the same set of samples from each data type, such as mRNA expression, DNA methylation, and microRNA expression. Then, it fuses these networks into one similarity network. The key idea of fusion is to update one network by utilizing two pieces of information: (a) the local affinity of the network and (b) the average similarity matrix of all the other networks. An iterative fusion process takes place, which increases the similarity between networks with each iteration until SNF achieves a final network by taking the average of all networks. In summary, SNF makes use of a network's local structure, integrating both common and complementary information across networks. SNF has been applied to identify cancer subtypes and predict survival (Wang et al., 2014).

More network integration methods have been applied in genomics research in addition to the ones reviewed here, although they are often application specific and differ substantially from one another. For a more substantial review, we refer the readers to the review paper of Wani and Raza (2018). In general, network integration methods offer a simple and straightforward solution whereby similar nodes (e.g., genes and proteins) across multiple networks are integrated by merging different types of edges from multiple networks. Although simple, they are less efficient when it comes to preserving the relationships across multiple networks, particularly when the networks are heterogeneous and do not share the same biological mechanism.

UNIQUE CHALLENGES OF MICROBIOME MULTI-OMICS NETWORK ANALYSIS

Microbiome data science is often challenged by various statistical properties of microbiome data, including its compositionality, heterogeneity, and sparsity. These properties impact how statistical methods are applied to microbiome data and require careful consideration to ensure appropriate analysis. In this section, we discuss these various properties and how they impact the application of the approaches described in Section "Review of Available Network-Based Procedures" to microbiome data, especially with respect to microbiome multi-omics data integration. Our hope is that this discussion helps readers identify opportunities to transform microbiome multi-omics network analysis.

Compositionality

One of the unique characteristics of microbiome data is its compositionality. Microbiome data are often presented as the abundances of different microbial taxa contained in a microbial community. However, microbiome data only carry information about the relative abundances of the taxa instead of their true abundances. This is because the total sequence count of all taxa for each sample, known as the sequencing depth of the sample, is an experimental technicality imposed by the sequencing instrument and bears no biological relevance.

Therefore, the abundance count of a taxon in a sample only reflects the relative abundance of the taxon compared against all other taxa, rather than the absolute count of molecules in the underlying community attributable to the taxon. As a result, these data exist under an arbitrary sum constraint and are thus referred to as compositional data. This feature is also visualized in **Figure 1A**.

When modeling compositional data, it is important to account for the fact that the sum is uninformative about (i.e., ancillary for) the parameters of interest, and therefore, it may be desirable to consider the conditional distribution of the data regarding the sequencing depths as pre-fixed quantities. For example, a common strategy to acknowledge compositionality of microbiome data is to convert the abundance count of each taxon into proportions or relative abundances that sum up to one for each sample. A consequence of the sum constraint is that the features will tend to be negatively correlated even if the underlying (unobserved) true abundances are independent.

The traditional marginal correlation analysis methods in Section "Marginal Correlation Analysis" such as Pearson's, Spearman's, and Kendall's correlations do not consider microbiome data compositionality. The key issue is that there exists a constraint on the correlations between one taxon and all other taxa due to the compositionality of the data, which can yield spurious inferences of interaction. For example, for any given taxon, its Pearson's correlation coefficients with the other taxa always sum up to −1, regardless of how this taxon interacts with the rest of the microbiome. Recently, new methods have been proposed to account for data compositionality when constructing microbial networks. For example, SparCC (Friedman and Alm, 2012) employs a log-ratio transformation for every pair of taxa being correlated to remove compositionality: the ratio of the abundances of two taxa is independent of which other taxa are included in the analysis, a property termed subcompositional coherence. SparCC also uses an iterative algorithm that identifies the pair of taxa with the strongest correlation in each step and terminates iterations when a relatively sparse network structure is obtained. More recently, CCLasso (Fang et al., 2015) and REBACCA (Ban et al., 2015) use global optimization procedures that estimate the correlation network of all species while imposing an explicit constraint caused by the compositionality of the data and a sparsity constraint on the network. While this approach is effective at controlling for data compositionality, these methods are only designed to reconstruct taxon-taxon interaction networks. To the best of our knowledge, we are unaware of approaches that consider compositionality when constructing microbiome multi-omics networks.

The compositionality of microbiome data has also been considered in methods based on graphical models (Section "Methods Based on Graphical Models"). Given that the major goal of graphical modeling is to infer microbial interactions through the estimation of the inverse covariance matrix between species, it is harder to correct for data compositionality as compared to marginal correlation analysis. The unique challenge here is that the sum constraint in compositional data induces linear dependency between features and thus gives rise to a degenerate covariance matrix, meaning that the inverse covariance matrix does not exist. To overcome this challenge, Kurtz et al. (2015) proposed a method called SPIEC-EASI that first converts raw counts into relative abundances, i.e., the proportions of each taxon's abundance within a sample, and then uses the centered log-ratio transformation on the relative abundances. They further argue that the covariance matrix of the transformed relative abundances is a good approximation to that of the log-transformed raw counts. SPIEC-EASI uses both neighborhood selection (Meinshausen and Bühlmann, 2006) and graphical lasso (Friedman et al., 2008) to infer a sparse inverse covariance matrix for a network. In addition, Yang et al. (2017) proposed a method called mLDM that uses a hierarchical Bayesian model (lognormal-Dirichlet-multinomial) on the compositional counts and then estimates a sparse inverse covariance matrix between the species through maximizing the L1 penalized posterior distribution.

Compositionality is also important to consider in regression-based methods. In Section "Review of Available Network-Based Procedures", we reviewed several regression methods to construct biological networks. To apply these methods to integrate microbiome data and another data type, it is possible to use microbiome data as either predictors or responses. Therefore, we discuss these two situations separately. In the case that microbiome data are used as predictors, there are two major challenges: the high dimensionality of the data and a sum constraint on the predictors imposed by the compositional nature of the data. Lin et al. (2014) proposed an L1 regularization method for the linear log-contrast model that meets these unique challenges of compositional data to study the association between the microbial compositions and the response variable. Moreover, Shi et al. (2016) extended the previous method to consider the subcompositions of taxa, i.e., the composition of taxa that belong to a given higher level taxonomic rank, and studied whether the observed subcompositions are associated with the response variable. On the other hand, if microbiome data are used as responses, it is essential to incorporate an appropriate distribution in the model to reflect the compositionality. For example, Chen and Li (2013) applied the Dirichlet-multinomial regression to investigate the association between microbiome composition and environmental covariates. Furthermore, Xia et al. (2013) proposed to use the logistic normal multinomial regression model to link covariates with taxonomic counts, given that the logistic normal distribution has a more flexible covariance structure than the Dirichlet distribution. The mLDM method (Yang et al., 2017) also investigates the association between the taxonomic counts and the environmental factors in their lognormal-Dirichlet-multinomial model.

As mentioned above, many network analysis methods have been proposed to consider the compositionality of the microbiome data. However, very few of them have been applied for network analyses that integrate multi-omics data alongside microbiome measures. We anticipate that this will be an active research area in the near future. Moreover, technological developments in microbiome data science, including the estimation of absolute cellular abundances from microbiome sequence data (Vandeputte et al., 2017) may help offset the need to correct for data compositionality when reconstructing microbiome networks.

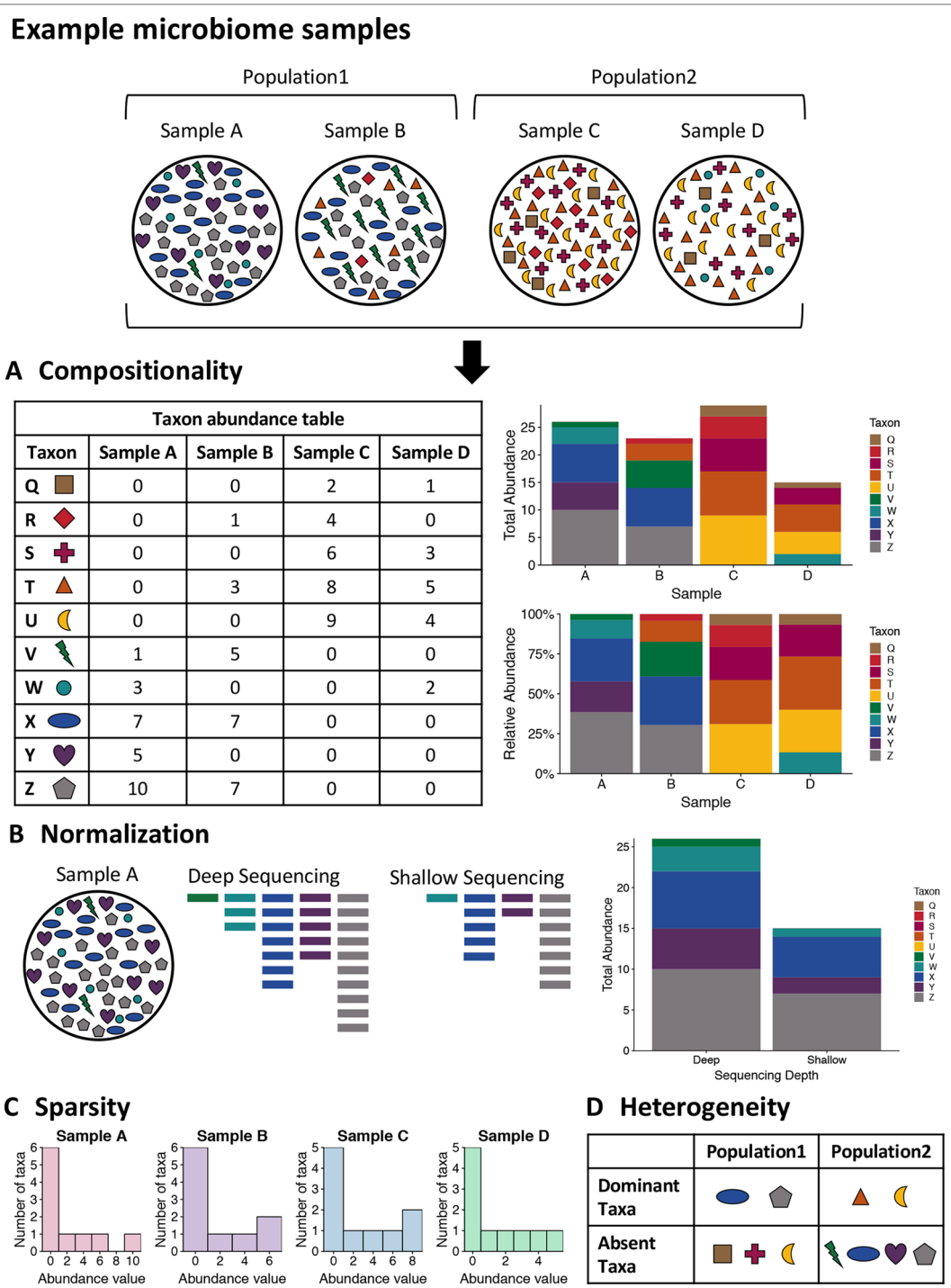


FIGURE 1 | Visualizing the unique challenges of microbiome data. A mock set of bacterial samples from two populations where each colored shape is a bacterial taxon. **(A)** Compositionality. The taxon abundance table depicts the count of each observed taxon in each sample. When sequencing microbiome samples, the resulting counts of taxa are not representative of the actual taxa counts in the sample due to constraints of sequencing. Due to this, relative abundances are generally used in analysis of microbiome data. The bar plots illustrate the difference in community representation between raw counts (top) and relative abundances (bottom). **(B)** Normalization. Due to the constraints of sequencing, the overall sequencing depth of a sample can impact the results. For example, shallow sequencing may miss rare taxa such as the green taxon V in the example sample A that is present in low abundance in the community. **(C)** Sparsity. Microbiome data are often very sparse, where most observations are zero. This is illustrated by the histogram of taxa counts for each sample where most counts are zero and there are few taxa with high counts. This can also be seen in the table for part A, where many entries are zero. **(D)** Heterogeneity. The table summarizes the taxonomic heterogeneity in the mock dataset between the two populations. Each sample has a unique taxonomic composition, but there are also population specific signatures. The samples in each population are dominated by a few taxa, and these dominant taxa are different for the two populations. Additionally, there are taxa that are highly abundant in one sample and absent from the rest, such as the purple taxon Y in sample A.

Normalization

Similar to many other omics data, microbiome data can exhibit strong heterogeneity from one study to another or from one biological sample to another even in the same study. For example, microbiome data may be collected from different geographic populations and they may have very different taxonomic distributions (He et al., 2018). In addition, varying data generation and processing procedures for microbiome data can also lead to heterogeneity across studies. For example, different sequencing technologies will result in different sequence lengths across studies, which can impact the discovery of taxa. Moreover, different studies may apply different data processing procedures (e.g., how sequences are assigned to taxonomic units or phylotypes) that may impact the distribution of taxa across studies.

One unique heterogeneity between studies or between samples in microbiome data is the variation of sequencing depths, as visualized in **Figure 1B**. Sequencing depth, the total count of sequences generated across all taxa for a biological sample, is an experimental technicality and often varies considerably across samples in a microbiome sequencing experiment. Like other omics data, normalization is an important and often first analytical step. The traditional approaches for normalizing microbiome data is either to transform count-based measures of taxa into relative abundances (i.e., proportions) of the taxa or to rarefy the counts, i.e., subsampling without replacement from each sample such that all samples have the same number of total counts across taxa. In addition, alternative normalization methods using other criteria are also used in the microbiome research community, including upper quantile normalization (Bullard et al., 2010), CSS normalization (Paulson et al., 2013), variance stabilizing transformation (Love et al., 2014), and trimmed mean of M-values normalization (Robinson et al., 2009; McCarthy et al., 2012). Most of these alternative normalization methods are borrowed from the techniques for RNA-seq data analysis. While these alternative methods are advocated in studies that focused on differential abundance testing, the traditional approaches of proportion- and rarefaction-based normalization provide more accurate community-level comparisons (McKnight et al., 2018).

Studies have also assessed the influence of sequencing depth on the quality of microbiome data. For example, Jovel et al. (2016) measured the minimum sequencing depth that can still provide a consistent taxonomic classification by randomly sampling from a sequencing library with different depths, while Nayfach et al. (2015) conducted a similar analysis for the functional annotation of metagenomes. Zaheer et al. (2018) evaluated the impact of sequencing depth on the characterization of the microbiome and resistome and indicated that the relative proportions of sequence assignments remained fairly constant regardless of depth. Although these studies show that taxonomic and functional annotation is fairly stable regardless of the sequencing depth, McMurdie and Holmes (2014) argued that current practice in the normalization of microbiome count data is inefficient in the statistical sense. One key issue with rarefaction is that while it maintains the mean of the taxonomic proportions it ignores the variation of the proportions. For example, two equal proportions of an OTU in two samples can have unequal variances due to the different sequencing depths between the two samples. This problem of unequal variances is called "heteroscedasticity" and is not accounted for during typical rarefaction approaches. Heteroscedasticity could impact downstream analysis such as differential abundance analysis and construction of microbial networks.

In Section "Compositionality", we reviewed statistical models such as Dirichlet-multinomial regression (Chen and Li, 2013), logistic normal multinomial regression (Xia et al., 2013), and mLDM (Yang et al., 2017). These models not only consider the compositionality of microbiome data but also take the heteroscedasticity into account because the sequencing depth is explicitly modeled in the multinomial distribution. However, most of the above methods are applied to identify the association between the taxonomic composition and the environmental factors. While these models are potentially applicable to network analyses that integrate microbiome and other omics data, further investigations are warranted, especially considering the scale of the dimensionality of multi-omics data.

Sparsity

Taxonomic abundance data are typically sparse in nature, meaning that a high proportion of the counts are zeros (Paulson et al., 2013). This feature of microbiome data frequently poses challenges to common statistical methods, and tailored techniques are often required to properly analyze microbiome data and to integrate them with other omics data. For example, due to the compositionality of microbiome data (see Section "Compositionality"), many statistical methods utilize transformations that involve taking logarithms on the counts or ratios between them. However, zero counts cause a technical problem for these transformations. To circumvent this issue, a widely used strategy is to add a small constant to all count measures, known as a pseudo-count (Kurtz et al., 2015; Mandal et al., 2015), or to replace the zeros by an estimated value (Palarea-Albaladejo and Martín-Fernández, 2015; Gloor et al., 2016). Some recent work has studied the problem of how to best choose the pseudo-count and how to find the estimated value (Martín-Fernández et al., 2003; Martín-Ferńandez et al., 2011; Martín-Fernández et al., 2015). However, more research is needed to determine how these techniques impact integrative network estimation for microbiome multi-omics data.

The sparsity of microbiome data also challenges modeling. The excess zeros, coupled with a high frequency of a very low number of observations per taxon, results in a heavily skewed distribution of taxon counts across samples, with a large point mass at zero and a long right tail. This is also visualized *via* a mock dataset in **Figure 1C**. Consequently, network estimation methods that work well for continuous data, including those assuming that the counts follow a Gaussian distribution such as graphical lasso, may not work well when directly applied to such data because of poor model fit. Nonparametric correlation measures such as Spearman's rank correlation and Kendall's tau can be used to avoid an assumption of normality and tackle highly skewed data. However, the power of such methods may deteriorate when data measures distribute with a point mass at

zero, as this mass of zeros leads to a large number of ties that complicate rank-based measures of correlation (Huson, 2007). In addition, agglomeration of taxon measures into higher order taxonomic groups may reduce the effects of sparsity and improve alignment between the observed data distributions and model assumptions. However, such agglomerative procedures can erode resolution of specific taxonomic units that manifest important and nuanced relationships with other study covariates.

In recent years, a variety of probability models have been developed for microbiome count data. The Poisson or negative binomial distributions have been useful for analyzing count data from other types of sequencing studies, such as transcriptomic studies using RNA sequencing. However, microbiome data often—though not always—exhibit more zeros and heavier skewness than expected from these models. To this end, zero-inflated models (Sharpton et al., 2017) and hurdle models (Hu et al., 2011) have been proposed. For example, the zero-inflated Poisson distribution considers a mixture of a Poisson distribution and a probability mass at zero to account for the large frequency of zeros in microbiome data (Xu et al., 2015). However, most of these methods focus on modeling the marginal distribution of a single taxon at a time and are not directly applicable to the joint modeling of multiple taxa and therefore cannot be used for microbial network estimation.

Another type of models used for microbiome count data is the Dirichlet-multinomial model and its zero-inflated versions. It has been used in a number of methods to model the multivariate distribution of the counts of a collection of taxa (Holmes et al., 2012; Chen and Li, 2013; Tang and Chen, 2018). However, a criticism of these methods is that the Dirichlet-multinomial distribution imposes a negative correlation between the abundances of any given pair of taxa. This inflexibility in the correlation structure makes such methods particularly problematic when used to infer the interaction between taxa. A promising approach to addressing this pitfall is to consider a hierarchical model where the conditional distribution of the observed counts is modeled by a multivariate count distribution such as multinomial distribution or Dirichlet-multinomial distribution, whose parameters are linked to a multivariate continuous distribution, such as multivariate normal distribution, that allows a flexible and realistic correlation structure (Xia et al., 2013; Yang et al., 2017).

Despite the success of the aforementioned models for microbiome count data, their use has for the most part been limited to differential abundance analysis, where the abundance of individual or groups of taxa is associated with an environmental factor of interest. Further work is needed to explore their applicability to multi-omics data and integrative network analysis. We see it as a great research opportunity to combine these models with cutting edge multi-omics network estimation methods to make the latter more appropriate for microbiome studies.

Heterogeneity

Related to the issue of sparsity is the heterogeneity exhibited in studies that survey the composition of microbial communities. The composition of microbial communities often varies tremendously across hosts and environments. For example, it is not uncommon to observe that a taxon that is relatively abundant in one person's gut while being completely absent in another's; for a given taxon, it is often the case that only a proportion of the samples have nonzero abundance. While the number of observed taxa from the entire data set may be large, the microbiota in any given sample tend to be dominated by only a relatively small number of taxa with high abundance, with the rest of the taxa having zero and very low counts. Moreover, the set of dominant taxa can vary drastically from individual to individual. We call the above phenomena taxonomic heterogeneity, as visualized in **Figure 1D**. It results in a unique characteristic of microbiome data sets that features (i.e., taxa) present in all samples are rare and those present in a small proportion of samples prevail. This is in contrast to most other types of omics data such as transcriptomic data, where the majority of genes are expected to have nonzero expression levels in all samples.

Different approaches have been applied to account for taxonomic heterogeneity when measuring the interaction between two microbial taxa or between a taxon and another biological feature (e.g., a metabolite). The most commonly used strategy is to include the data from all biological samples, regardless of whether the taxon of interest is present or not. An alternative strategy is to exclude the samples in which the given taxon is not present and only consider those abundance data that are nonzero for the taxon. A third strategy focuses on the dichotomous outcome of whether a taxon is present or absent in individual samples, while ignoring the actual abundance (Mainali et al., 2017; Albayrak et al., 2018). The first approach regards a sample where a taxon is absent as having "zero abundance" of the taxon, which is only quantitatively, but not qualitatively, different from a sample where the abundance of the taxon is very low. This approach's main advantage is that no information is discarded from the data, whereas the latter two approaches each discard part of the data. Most methods using the first approach assume that, if a biological interaction exists between a microbial taxon T and another feature M (e.g., a metabolite), the feature M is associated with the abundance of T in the same way that it is associated with the occurrence of T in a community. However, the biological process in which M is involved in the introduction or establishment of T may in theory be very different from the one in which M impacts its abundance. For example, M may promote the growth of T in a person's gut microbiome only if it already contains T. It is also possible that elevated levels of M are associated with increasing a person's chance of exposure to T and consequently its presence in the gut, but do not affect its abundance. For these types of relationships, the latter two strategies may have merits.

In addition to taxonomic heterogeneity, functional heterogeneity is another feature of microbiome data that challenges statistical methods for network inference. Most current methods for microbial network estimation, such as those by Kurtz et al. (2015) and Yang et al. (2017), assume that there exists a common microbial network underlying all samples in the data. However, the interaction between two microbial taxa or between a taxon and another type of feature may be context

dependent and may vary from sample to sample. For example, the interaction between taxa in the human gut may depend upon the enterotypic context of the individual's gut microbiome. Recent statistical developments have been made on the joint estimation of multiple graphical models, which assumes the samples are from several known subpopulations (e.g., corresponding to several biological conditions) and allows a different network to be inferred for each subgroup (Chun et al., 2015; Lin et al., 2017). In addition, some emergent methods have been applied to genomic data to allow network heterogeneity among all samples, between or within biological conditions. For example, Luo and Wei (2018) developed a nonparametric Bayesian method to estimate dynamic transcription factor networks by borrowing information across biological conditions and meanwhile allowing heterogeneity across samples. Another example is mixGlasso (Städler et al., 2017), a latent variable extension of graphical lasso, which uses a mixture model to allow samples to be clustered into groups that can have different networks. Despite these recent statistical developments, methods have not been established to address the unique needs of microbiome data analysis and for the purpose of integrating microbiome multi-omics data.

DISCUSSION

This review focuses on statistical network analysis methods that have been applied or have great potential to be applied to multi-omics integration of microbiome data. Therefore, this review does not cover some of the other analytical methods and tools that are either not directly relevant to statistical network analysis or not specific to microbiome data but are still applicable to general multi-omics integration. For these more general methods and tools, we refer the readers to the following review papers. Bersanelli et al. (2016) categorized various data integration methods into four classes according to whether they are Bayesian and whether they are network-based, and they reviewed each class of methods focusing on their mathematical and methodological aspects. Li et al. (2018) provided a comprehensive review on omics and clinical data integration techniques from a machine learning perspective. Huang et al. (2017) separately reviewed unsupervised, supervised, and semisupervised data integration tools and their applications to predicting patient survival. Zeng and Lumley (2018) reviewed the traditional statistical methods of exploratory and supervised learning as well as their variations tailored to multi-omics studies. Mirza et al. (2019) discussed state-of-the-art machine learning-based approaches for tackling five specific computational challenges associated with integrative analysis: curse of dimensionality, data heterogeneity, missing data, class imbalance, and scalability issues.

While our review focuses on data analysis, it is important to note that study design and data collection can impact data integration-based investigations. For example, in a multi-omics study, it is rarely the case that researchers are able to collect a complete data set in the sense that all feature types are measured for all samples. This incomplete coverage of samples can dramatically reduce the set of samples subject to integration. In a longitudinal multi-omics study of the gut microbial ecosystem in inflammatory bowel diseases (Lloyd-Price et al., 2009), 132 participants were followed for one year and their stool samples were collected every two weeks, resulting in 1,785 stool samples. However, given the difficulty of collecting all feature types (for example, metagenomics, metatranscriptomics, proteomics, metabolomics, etc.) at each timepoint, the final data include only 305 samples that yielded all stool-derived feature types, whereas 791 samples offered paired metagenomic and metatranscriptomic data. As exemplified in this study, to derive networks depicting the relationships between certain pairs of feature types, one may need to rely on separate sets of samples for the two feature types. This strategy, compared with one in which paired multi-omic data are available on a common set of samples, would impact the accuracy and interpretation of the resulting networks. In addition to the above practical issue of missing data, considerations of study design can impact integration, such as whether the samples were collected longitudinally or cross-sectionally. Given that this is a very broad topic, we refer readers to additional review papers (Franzosa et al., 2015; Buescher and Driggers, 2016; Haas et al., 2017; Hasin et al., 2017) for more detailed discussions about how study design impacts multi-omics investigations.

The recent work by the Integrative Human Microbiome Project (iHMP, https://hmpdacc.org/ihmp/) exemplifies the power and promise of microbiome multi-omic data integration. As the second phase of the NIH Human Microbiome Project, iHMP aimed to link interactions between humans and their microbiomes to health-related outcomes by analyzing data sets on microbiome and host activities in longitudinal studies of disease-specific cohorts (Integrative HMP (iHMP) Research Network Consortium 2014; Integrative HMP (iHMP) Research Network Consortium 2019). Fortunately for the research community, the iHMP has made these measures publicly available as downloadable datasets that can serve as resources to test and evaluate new models, methods, and analyses, including the network methods reviewed in this paper. In fact, many of the individual studies conducted as part of iHMP have applied and/or developed network-based methods for integrating multi-omics data. For example, Lloyd-Price et al. (2019) applied integrative analysis to identify microbial, biochemical, and host factors central to the functional dysbiosis in the gut microbiome during inflammatory bowel disease activity. They constructed networks for associations of features from 10 feature types: metagenomic species, species-level transcription ratios, functional profiles at the Enzyme Commission level (metagenomes, metatranscriptomes, and proteomes), metabolites, host transcription (rectal and ileal separately), serology, and fecal calprotectin. In particular, they used mixed-effects regression models (which belong to the regression-based methods discussed in Section "Regression-Based Methods") to remove subject-specific random effects and covariate effects from each feature type, and then applied Spearman correlation (which belong to the marginal correlation analysis methods discussed in Section "Marginal Correlation Analysis") to the resulting residuals to construct cross-feature type interactions.

We conclude this review with some final thoughts about microbiome multi-omics network analysis. Integrative network analysis holds great potential to resolve how microbes interact among themselves and with their environment. However, the

application of such analyses to microbiome data remains nascent, and the requisite analytical tools have only begun to emerge. Fortunately, a growing number of statistical methods have been developed in the fields of network estimation and multi-omics data analysis, which provide a promising pool of ideas and methodologies to potentially borrow from. However, when applying these existing tools to microbiome multi-omics network inference, it is important to consider the limitations of the underlying methodologies and their applicability to microbiome studies. In particular, the unique features of microbiome data present pressing statistical challenges and often call for tailored computational tools. A thorough understanding of the unmet statistical needs and specific properties of microbiome data is critical to the innovation of efficient, robust, and scalable network inference methodologies suitable for microbiome multi-omics network inference. Meanwhile, awareness of the analytical challenges associated with microbiome data can facilitate the development of new study designs and technologies that have the potential to mediate some of the major limitations currently hindering microbiome data analytics. An emerging example is the coupling of 16S data with measures of the total abundance of microorganisms in a sample, which is a possible way of circumventing the compositionality constraint in microbiome data. Going forward, joint statistical, scientific, and technological efforts will help promote the application of multi-omics network analysis to solve pressing problems in microbiome science.

AUTHOR CONTRIBUTIONS

DJ, TS, and YJ led and conducted the review. CA, CH, MM, and CT contributed equally to the review and wrote the first draft of sections of the manuscript. DJ, TS, and YJ wrote the first draft of sections of the manuscript and contributed to the manuscript revision. All authors read and approved the final version.

FUNDING

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM126549. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell.* 143, 1005–1017. doi: org/10.1016/j.cell.2010.11.013

Albayrak, L., Khanipov, K., Golovko, G., and Fofanov, Y. (2018). Detection of multi-dimensional co-exclusion patterns in microbial communities. *Bioinformatics (Oxford, England).* 34, 3695–3701. doi: org/10.1093/bioinformatics/bty414

Alivisatos, A. P., Blaser, M. J., Brodie, E. L., Chun, M., Dangl, J. L., Donohue, T. J., et al. (2015). A unified initiative to harness Earth's microbiomes. *Science*. 350, 507–508. doi: org/10.1126/science.aac8480

Amano, S. I., Ogawa, K. I., and Miyake, Y. (2018). Node property of weighted networks considering connectability to nodes within two degrees of separation. *Sci. Rep.* 8, 8464. doi: org/10.1038/s41598-018-26781-y

Aylward, F. O., Eppley, J. M., Smith, J. M., Chavez, F. P., Scholin, C. A., and DeLong, E. F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl Acad. Sci.* 112, 5443–5448. doi: org/10.1073/pnas.1502883112

Bakker, O. B., Aguirre-Gamboa, R., Sanna, S., Oosting, M., Smeekens, S. P., Jaeger, M., et al. (2018). Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nat. Immunol.* 19 (7), 776–786. doi: org/10.1038/s41590-018-0121-3

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*. 31, 3322–3329. doi: org/10.1093/bioinformatics/btv364

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2003). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci.* 101, 3747–3752. doi: org/10.1073/pnas.0400087101

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 17 (2), 15. doi: org/10.1186/s12859-015-0857-9

Bickel, P. J., and Li, B. (2006). Regularization in statistics. *Test*. 15, 271–344. doi: org/10.1007/BF02607055

Blaser, M. J., Cardon, Z. G., Cho, M. K., Dangl, J. L., Donohue, T. J., Green, J. L., et al. (2016). Toward a predictive understanding of earth's microbiomes to address 21st century challenges. *MBio*. doi: 10.1128/mbio.00714-16

Bouslimani, A., Porto, C., Rath, C. M., Wang, M., Guo, Y., Gonzalez, A., et al. (2015). Molecular cartography of the human skin surface in 3D. *Proc. Natl Acad. Sci.* 112, E2120–E2129. doi: 10.1073/pnas.1424409112

Buescher, J. M., and Driggers, E. M. (2016). Integration of omics: More than the sum of its parts. *Cancer Metab.* 4, 4. doi: 10.1186/s40170-016-0143-y

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94. doi: 10.1186/1471-2105-11-94

Burges, C. J. C. (2009). Dimension Reduction: A Guided Tour. *Found. Trends® Mach. Learn.* 2, 275–364. doi: 10.1561/2200000002.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* 100, 139–156. doi: 10.1093/biomet/ass058

Chaibub-Neto, E., Keller, M. P., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes, Supplement. *Ann. Appl. Stat.* 4 (1), 320–339. doi: 10.1214/09-AOAS288SUPP

Charitou, T., Bryan, K., and Lynn, D. J. (2016). Using biological networks to integrate, visualize and analyze genomics data. *Genet. Sel. Evol.* 48, 27. doi: 10.1186/s12711-016-0205-1

Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. App. Stat.* 7, 418–442. doi: 10.1214/12-AOAS592

Cho, H., Berger, B., and Peng, J. (2015). Diffusion component analysis: unraveling functional topology in biological networks. *Research in Computational Molecular Biology, Lecture Notes in Computer Science, 9029* (Springer, Cham.), 62–64. doi: 10.1007/978-3-319-16706-0_9

Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* 3 (6), 540–548.e5. doi: 10.1016/j.cels.2016.10.017

Chun, H., Chen, M., Li, B., and Zhao, H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Front. Genet.* 4, 294. doi: 10.3389/fgene.2013.00294

Chun, H., and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat.* 72, 3–25. doi: 10.1111/j.1467-9868.2009.00723.x

Chun, H., Zhang, X., and Zhao, H. (2015). Gene regulation network inference with joint sparse gaussian graphical models. *J. Comput. Graph. Stat.* 24, 954–974. doi: 10.1080/10618600.2014.956876

Chung, D., and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Stat. App. Genet. Mol. Biol.* 9. doi: 10.2202/1544-6115.1492

Cranmer, S. J., Leifeld, P., McClurg, S. D., and Rolfe, M. (2017). Navigating the Range of Statistical Tools for Inferential Network Analysis. *Am. J. Pol. Sci.* 61, 237–251. doi: 10.1111/ajps.12263

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A. K., Sempoux, C., et al. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Med.* 1, 39. doi: 10.1186/gm39

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B Stat.* 76, 373–397. doi: 10.1111/rssb.12033

Dao, M. C., Sokolovska, N., Brazeilles, R., Affeldt, S., Pelloux, V., Prifti, E., et al. (2019). A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity. *Front. Physiol.* 9. doi: 10.3389/fphys.2018.01958

Dohlman, A. B., and Shen, X. (2019). Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. *Exp. Biol. Med.* 244, 445–458. doi: 10.1177/1535370219836771

Dorogovtsev, S. N., and Mendes, J. F. F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford: Oxford University Press. doi: 10.1063/1.1825279

Drton, M., and Maathuis, M. H. (2017). Structure Learning in Graphical Modeling. *Annu. Rev. Stat. Its Appl.* 4, 365–393. doi: 10.1146/annurev-statistics-060116-053803

Engel, D., Hüttenberger, L., and Hamann, B. (2011). A survey of dimension reduction methods for high-dimensional data analysis and visualization. *oasics-OpenAccess Ser. Inf.* 27, 135–149. doi: 10.4230/OASIcs.VLUDS.2011.135

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*. 31, 3172–3180. doi: 10.1093/bioinformatics/btv349

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

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W. (2007). Emergent biogeography of microbial communities in a model ocean. *Science*. 315, 1843–1846. doi: 10.1126/science.1138544

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular "omics" for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 9, 432–441. doi: 10.1093/biostatistics/kxm045

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

Frost, H. R., and Amos, C. I. (2018). A multi-omics approach for identifying important pathways and genes in human cancer. *BMC Bioinformatics*. 19, 479. doi: 10.1186/s12859-018-2476-8

Fujita, A., Vidal, M. C., and Takahashi, D. Y. (2017). A statistical method to distinguish functional brain networks. *Front. Neurosci.* 11. doi: 10.3389/fnins.2017.00066

Furlotte, N. A., Kang, H. M., Ye, C., and Eskin, E. (2011). Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics*. 27, i288–i294. doi: 10.1093/bioinformatics/btr221

Gade, S., Porzelius, C., Fälth, M., Brase, J. C., Wuttig, D., Kuner, R., et al. (2011). Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. *BMC Bioinformatics* 12 (1), 488. doi: 10.1186/1471-2105-12-488

Gao, B., and Cui, Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics*. 31, 3953–3960 doi: 10.1093/bioinformatics/btv513

Gaulke, C. A., Barton, C. L., Proffitt, S., Tanguay, R. L., and Sharpton, T. J. (2016). Triclosan exposure is associated with rapid restructuring of the microbiome in adult zebrafish. *PLoS One* 11, e0154632. doi: 10.1371/journal.pone.0154632

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261. doi: 10.1093/nar/gkh036

Gloor, G. B., Macklaim, J. M., Vu, M., and Fernandes, A. D. (2016). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Aust J. Stat.* 45, 73-87. doi: 10.17713/ajs.v45i4.122

Goh, G., Dey, D. K., and Chen, K. (2017). Bayesian sparse reduced rank multivariate regression. *J. Multivariate Anal.* 157, 14–28. doi: 10.1016/j.jmva.2017.02.007

Gould, A. L., Zhang, V., Lamberti, L., Jones, E. W., Obadia, B., Korasidis, N., et al. (2018). Microbiome interactions shape host fitness. *Proc. Natl Acad. Sci.* 115, E11951–E11960. doi: 10.1073/pnas.1809349115

Griffiths-Jones, S., Saini, H. K., Van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36, D154–D158. doi: 10.1093/nar/gkm952

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*. 98, 1–15. doi: 10.1093/biomet/asq060

Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S. J., and Ralser, M. (2017). Designing and interpreting "multi-omic" experiments that may change our understanding of biology. *Curr. Opin. Syst. Biol.* 6, 37–45. doi: 10.1016/j.coisb.2017.08.009

Hardoon, D. R., and Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* 83 (3), 331–353. doi: 10.1007/s10994-010-5222-7

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18 (1), 1–15. doi: 10.1186/s13059-017-1215-1

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

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 (1), 1–12. doi: 10.1038/nmicrobiol.2016.180

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 12, 55–67. doi: 10.1080/00401706.1970.10488634

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

Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.* 41 (8), e95. doi: 10.1093/nar/gkt145

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P. M., van Eijk, K., et al. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 13, R97. doi: 10.1186/gb-2012-13-10-r97

Hu, M. C., Pavlicova, M., and Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *Am. J. Drug Alcohol Abuse* 37, 367–375. doi: 10.3109/00952990.2011.597280

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8, 1–12. doi: 10.3389/fgene.2017.00084

Huson, L. W. (2007). Performance of some correlation coefficients when applied to zero-clustered data. *J. Mod. Appl. Stat. Methods*. 6, 530–536. doi: 10.22237/jmasm/1193890560

Isci, S., Dogan, H., Ozturk, C., and Otu, H. H. (2014). Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics*. 30, 860–867. doi: 10.1093/bioinformatics/btt643

Jovanović, I., Živković, M., Jovanović, J., Djurić, T., and Stanković, A. (2014). The co-inertia approach in identification of specific microRNA in early and advanced atherosclerosis plaque. *Med. Hypotheses*. 83, 11–15. doi: 10.1016/j.mehy.2014.04.019

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* 7, 459. doi: 10.3389/fmicb.2016.00459

Kadarmideen, H. N., Watson-Haigh, N. S., and Andronicos, N. M. (2011). Systems biology of ovine intestinal parasite resistance: Disease gene modules and biomarkers. *Mol. BioSyst.* 7, 235–246. doi: 10.1039/c0mb00190b

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Fröhlich, H. (2018). Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. Rep.* 8 (1), 1–13. doi: 10.1038/s41598-018-29433-3

Kim, D. C., Kang, M., Zhang, B., Wu, X., Liu, C., and Gao, J. (2014). "Integration of DNA methylation, copy number variation, and gene expression for gene regulatory network inference and application to psychiatric disorders". *In proceedings–IEEE 14th International Conference on Bioinformics and Bioengineering, BIBE 2014*, 238–242. doi: 10.1109/BIBE.2014.71

Kim, S. H., Jhong, J. H., Lee, J. J., Koo, J. Y., Lee, B. Y., and Han, S. W. (2017). Node-structured integrative gaussian graphical model guided by pathway information. *Comput. Math. Methods Med.* 1–10. doi: 10.1155/2017/8520480

Kim, S., Sohn, K. A., and Xing, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* 25, i204–i212. doi: 10.1093/bioinformatics/btp218

Kint, G., Fierro, C., Marchal, K., Vanderleyden, J., and De Keersmaecker, S. C. J. (2010). Integration of 'omics data: does it lead to new insights into host-microbe interactions? *Future Microbiol.* 5, 313–328. doi: 10.2217/fmb.10.1

Kleaveland, B., Shi, C. Y., Stefano, J., and Bartel, D. P. (2018). A network of noncoding regulatory RNAs acts in the mammalian brain. *Cell*. 174, 350-362. e17. doi: 10.1016/j.cell.2018.05.022

Korb, K. B., and Nicholson, A. E. (2008). The causal interpretation of Bayesian networks. *Stud. Comput. Intell.* 83–116. doi: 10.1007/978-3-540-85066-3_4

Koski, T. J. T., and Noble, J. (2014). A review of bayesian networks and structure learning. *Math. Applicanda*. 40. doi: 10.14708/ma.v40i1.278

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

Lai, P. L., and Fyfe, C. (2003). KERNEL and nonlinear canonical correlation analysis. *Int. J. Neural Syst.* 10, 365–377. doi: 10.1142/s012906570000034x

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

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

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Stat. App. Genet. Mol. Biol.* 7. doi: 10.2202/1544-6115.1390

Lecca, P., and Re, A. (2015). Detecting modules in biological networks by edge weight clustering and entropy significance. *Front. Genet.* 6, 265. doi: 10.3389/fgene.2015.00265

Lee, D., Lee, W., Lee, Y., and Pawitan, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemom. Intell. Lab. Syst.* 109, 1–8. doi: 10.1016/j.chemolab.2011.07.002

Li, B., Chun, H., and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Am. Stat. Assoc.* 107, 152–167. doi: 10.1080/01621459.2011.644498

Li, H., Wang, Y., Jiang, J., Zhao, H., Feng, X., Zhao, B., et al. (2019). A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.* 10, 676. doi: 10.3389/fmicb.2019.00676

Li, W., Liu, C. C., Zhang, T., Li, H., Waterman, M. S., and Zhou, X. J. (2011). Integrative analysis of many weighted Co-Expression networks using tensor computation. *PLoS Comput. Biol.* 7, e1001106. doi: 10.1371/journal.pcbi.1001106

Li, Y., Wu, F. X., and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Brief Bioinf.* 19 (2), 325–340. doi: 10.1093/bib/bbw113

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*. 101, 785–797. doi: 10.1093/biomet/asu031

Lin, Z., Wang, T., Yang, C., and Zhao, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* 73 (3), 769–779.

Liu, Y., Liu, A., Liu, X., and Huang, X. (2019). A statistical approach to participant selection in location-based social networks for offline event marketing. *Information Sci.* 480, 90–108. doi: 10.1016/j.ins.2018.12.028

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

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

Luo, X., and Wei, Y. (2018). Nonparametric bayesian learning of heterogeneous dynamic transcription factor networks. *Ann. Appl. Stat.* 12 (3), 1749–1772. doi: 10.1214/17-AOAS1129

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8 (1), 573. doi: 10.1038/s41467-017-00680-8

Ma'ayan, A. (2011). Introduction to network analysis in systems biology. *Sci. Signaling* 4, tr5. doi: 10.1126/scisignal.2001965

Maier, T. V., Lucio, M., Lee, L. H., VerBerkmoes, N. C., Brislawn, C. J., Bernhardt, J., et al. (2017). Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. *MBio*. 8. doi: 10.1128/mbio.01343-17

Mainali, K. P., Bewick, S., Thielen, P., Mehoke, T., Breitwieser, F. P., Paudel, S., et al. (2017). Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. *PLoS One* 12, e0187132. doi: 10.1371/journal.pone.0187132

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., et al. (2018). Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci. Rep.* 8, 5875. doi: 10.1038/s41598-018-23931-0

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. doi: 10.3402/mehd.v26.27663

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* 35, 253–278. doi: 10.1023/A:1023866030544

Martín-Ferńandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). "Dealing with Zeros," in *Compositional Data Analysis: Theory and Applications*, 43–58. doi: 10.1002/9781119976462.ch4

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Modell.* 15 (2), 134–158. doi: 10.1177/1471082X14535524

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

McGrail, D. J., Federico, L., Li, Y., Dai, H., Lu, Y., Mills, G. B., et al. (2018). Multi-omics analysis reveals neoantigen-independent immune cell infiltration in copy-number driven cancers. *Nat. Commun.* 9, 1317. doi: 10.1038/s41467-018-03730-x

McHardy, I. H., Goudarzi, M., Tong, M., Ruegger, P. M., Schwager, E., Weger, J. R., et al. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 1, 17. doi: 10.1186/2049-2618-1-17

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. (2018). Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.* 10, 389–400. doi: 10.1111/2041-210X.13115

McMurdie, P. J., and Holmes, S. (2014). Waste Not, Want Not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10, e1003531. doi: 10.1371/journal.pcbi.1003531

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34, 1436–1462. doi: 10.1214/009053606000000281

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 15, 162. doi: 10.1186/1471-2105-15-162

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinf.* 17 (4), 628–641. doi: 10.1093/bib/bbv108

Min, E. J., Safo, S. E., and Long, Q. (2018). Penalized co-inertia analysis with applications to -omics data. *Bioinformatics* 35, 1018–1025. doi: 10.1093/bioinformatics/bty726

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes* 10, 87. doi: 10.3390/genes10020087

Mohammadnejad, A., Li, S., Duan, H., Lund, J., Li, W., Baumbach, J., et al. (2019). "Weighted gene co-expression network analysis of microarray mRNA expression profiling in response to electroacupuncture". *In proceedings–2018 IEEE International Conference on Bioinformics and Biomedicinem, BIBM 2018*. 1876–1883. doi: 10.1109/BIBM.2018.8621258

Moore, K. S., and t Hoen, P. A. C. (2019). Computational approaches for the analysis of RNA–protein interactions: a primer for biologists. *J. Biol. Chem.* 294, 1–9. doi: 10.1074/jbc.REV118.004842

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol.* 16, 67. doi: 10.1186/s13059-015-0637-x

Morgun, A., Dzutsev, A., Dong, X., Greer, R. L., Sexton, D. J., Ravel, J., et al. (2015). Uncovering effects of antibiotics on the host and microbiota using transkingdom gene networks. *Gut.* 64, 1732–1743. doi: 10.1136/gutjnl-2014-308820

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* 9, S4. doi: 10.1186/gb-2008-9-s1-s4

Nayfach, S., Bradley, P. H., Wyman, S. K., Laurent, T. J., Williams, A., Eisen, J. A., et al. (2015). Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Comput. Biol.* 11, e1004573. doi: 10.1371/journal.pcbi.1004573

Newman, M. (2010). "Networks: an introduction," in *Networks: An Introduction*. doi: 10.1093/acprof:oso/9780199206650.001.0001

Newman, M. E. J. (2004).). Analysis of weighted networks. *Phys. Rev. E: Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* 70, 056131. doi: 10.1103/PhysRevE.70.056131

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2014). Integrative Bayesian network analysis of genomic data. *Cancer Inf.* 13, 39–48. doi: 10.4137/CIn.s13786

Nie, L., Wu, G., Brockman, F. J., and Zhang, W. (2006a). Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics* 22 (13), 1641–1647. doi: 10.1093/bioinformatics/btl134

Nie, L., Wu, G., and Zhang, W. (2006b). Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: A multiple regression to identify sources of variations. *Biochem. Biophys. Res Commun.* 339, 603–610. doi: 10.1016/j.bbrc.2005.11.055

Opsahl, T., and Panzarasa, P. (2009). Clustering in weighted networks. *Soc. Networks* 31, 155–163. doi: 10.1016/j.socnet.2009.02.002

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.* 143, 85–96. doi: 10.1016/j.chemolab.2015.02.019

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.* 1, S119. doi: 10.1186/1753-6561-1-s1-s119

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. App. Genet. Mol. Biol.* 8, 1–34. doi: 10.2202/1544-6115.1406

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* 7. doi: 10.1093/gigascience/giy014

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., et al. (2012). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. App. Stat.* 4, 53–77. doi: 10.1214/09-AOAS271

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *J. Am. Stat. Assoc.* 110, 159–174. doi: 10.1080/01621459.2014.896806

Pfalzer, A. C., Kamanu, F. K., Parnell, L. D., Tai, A. K., Liu, Z., Mason, J. B., et al. (2016). Interactions between the colonic transcriptome, metabolome, and microbiome in mouse models of obesity-induced intestinal cancer. *Physiol. Genomics* 48, 545–553. doi: 10.1152/physiolgenomics.00034.2016

Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* 67 (3), 294–303. doi: 10.1016/j.ymeth.2014.03.006

Reverter, F., Vegas, E., and M., J. (2012). "Kernel methods for dimensionality reduction applied to the «omics» data," in *Principal component analysis - multidisciplinary applications*. 1–20. doi: 10.5772/37431

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rodrigues, R. R., Shulzhenko, N., and Morgun, A. (2018). Transkingdom networks: a systems biology approach to identify causal members of host–microbiota interactions. *Methods Mol. Biol.* 227–242. doi: 10.1007/978-1-4939-8728-3_15

Röttjers, L., and Faust, K. (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol. Rev.* 42, 761–780. doi: 10.1093/femsre/fuy030

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 22, 2532–2538. doi: 10.1093/bioinformatics/btl417

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). "Kernel principal component analysis BT - artificial neural networks — ICANN'97," in *Artificial Neural Networks — ICANN'97*.

Sharpton, T., Lyalina, S., Luong, J., Pham, J., Deal, E. M., Armour, C., et al. (2017). Development of inflammatory bowel disease is linked to a longitudinal restructuring of the gut metagenome in mice. *MSystems* 2. doi: 10.1128/mSystems.00036-17

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Ann. App. Stat.* 10, 1019–1040. doi: 10.1214/16-AOAS928

Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550. doi: 10.1038/ng.2982

Silk, M. J., Croft, D. P., Delahay, R. J., Hodgson, D. J., Weber, N., Boots, M., et al. (2017). The application of statistical network models in disease research. *Methods Ecol. Evol.* 8, 1026–1041. doi: 10.1111/2041-210X.12770

Städler, N., Dondelinger, F., Hill, S. M., Akbani, R., Lu, Y., Mills, G. B., et al. (2017). Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study. *Bioinformatics* 33, 2890–2896. doi: 10.1093/bioinformatics/btx322

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348, 1261359–1261359. doi: 10.1126/science.1261359

Suo, X., Minden, V., Nelson, B., Tibshirani, R., and Saunders, M. (2017). Sparse canonical correlation analysis. ArXiv Preprint ArXiv:1705.10865.

Tan, L., and Lei, D. (2013). Exact Solutions of a Generalized Weighted Scale Free Network. *J.Appl. Math.* 2013, 1–6. doi: 10.1155/2013/902519

Tang, Z.-Z., and Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, kxy025. doi: 10.1093/biostatistics/kxy025

Tap, J., Furet, J. P., Bensaada, M., Philippe, C., Roth, H., Rabot, S., et al. (2015). Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. *Environ. Microbiol.* 17, 4954–4964. doi: 10.1111/1462-2920.13006

Tapio, I., Fischer, D., Blasco, L., Tapio, M., Wallace, R. J., Bayat, A. R., et al. (2017). Taxon abundance, diversity, co-occurrence and network analysis of the ruminal microbiota in response to dietary changes in dairy cows. *PLoS One* 12, e0180260. doi: 10.1371/journal.pone.0180260

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K. A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15 (3), 569–583. doi: 10.1093/biostatistics/kxu001

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

The Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16, 276–289. doi: 10.1016/j.chom.2014.08.014

The Integrative HMP (iHMP) Research Network Consortium (2019). The integrative human microbiome project. *Nature*. doi: 10.1038/s41586-019-1238-8

Theriot, C. M., Bowman, A. A., and Young, V. B. (2016). Antibiotic-induced alterations of the gut microbiota alter secondary bile acid production and allow for clostridium difficile spore germination and outgrowth in the large intestine. *MSphere* 1. doi: 10.1128/mSphere.00045-15

Theriot, C. M., Koenigsknecht, M. J., Carlson, P. E., Hatton, G. E., Nelson, A. M., Li, B., et al. (2014). Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat. Commun.* 5, 3114. doi: 10.1038/ncomms4114

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. Ser. B Stat.* doi: 10.2307/2346178

Tong, M., Li, X., Parfrey, L. W., Roth, B., Ippoliti, A., Wei, B., et al. (2013). A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. *PLoS One* 8, e80702. doi: 10.1371/journal.pone.0080702

Vandeputte, D., Kathagen, G., D'Hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511. doi: 10.1038/nature24460

Waaijenborg, S., Verselewel De Witt Hamer, P. C., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat. App. Genet. Mol. Biol.* 7, 1–29. doi: 10.2202/1544-6115.1329

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi: 10.1038/nmeth.2810

Wang, Q., Wang, K., Wu, W., Giannoulatou, E., Ho, J. W. K., and Li, L. (2019). Host and microbiome multi-omics integration: applications and methodologies. *Biophys. Rev.* 11, 55–65. doi: 10.1007/s12551-018-0491-7

Wang, S., Cho, H., Zhai, C. X., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31, i357–i364. doi: 10.1093/bioinformatics/btv260

Wang, Y., Zhang, J., Xiao, X., Liu, H., Wang, F., Li, S., et al. (2016). The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. *Sci. Rep.* 6, 22722. doi: 10.1038/srep22722

Wani, N., and Raza, K. (2018). Integrative approaches to reconstruct regulatory networks from multi-omics data: a review of state-of-the-art methods. *Preprints* 1–20. doi: 10.20944/preprints201804.0352.v1

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534. doi: 10.1093/biostatistics/kxp008

Witten, D. M., and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. App. Genet. Mol. Biol.* 8, 1–27. doi: 10.2202/1544-6115.1470

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019). A selective review of multi-level omics data integration using variable selection. *High-Throughput* 8, 4. doi: 10.3390/ht8010004

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* 69, 1053–1063. doi: 10.1111/biom.12079

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One* 10, e0129606. doi: 10.1371/journal.pone.0129606

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593–597. doi: 10.1038/nature12364

Yang, X., Zhou, Y., Jin, R., and Chan, C. (2009). Reconstruct modular phenotype-specific gene networks by knowledge-driven matrix factorization. *Bioinformatics* 25 (17), 2236–2243. doi: 10.1093/bioinformatics/btp376

Yang, Y., Chen, N., and Chen, T. (2017). Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical bayesian statistical model. *Cell Syst.* 4, 129-137.e5. doi: 10.1016/j.cels.2016.12.012

Yuan, L., Guo, L. H., Yuan, C. A., Zhang, Y. H., Han, K., Nandi, A., et al. (2018). Integration of multi-omics data for gene regulatory network inference and application to breast cancer. *IEEE/ACM Transact. Comput. Biol. Bioinf* 16, 782–791. doi: 10.1109/TCBB.2018.2866836

Yuan, M., and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35. doi: 10.1093/biomet/asm018

Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S. R., Marinier, E., Van Domselaar, G., et al. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* 8, 5890. doi: 10.1038/s41598-018-24280-8

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining P-values. *Genet. Epidemiol.* 22, 170–1850. doi: 10.1002/gepi.0042

Zeng, I. S. L., and Lumley, T. (2018). Review of statistical learning methods in integrated omics studies (An integrated information science). *Bioinf. Biol. Insights* 12, 1–16. doi: 10.1177/1177932218759292

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. App. Genet. Mol. Biol.* 4, 17. doi: 10.2202/1544-6115.1128

Zhang, W., Bojorquez-Gomez, A., Velez, D. O., Xu, G., Sanchez, K. S., Shen, J. P., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* 50, 613–620. doi: 10.1038/s41588-018-0091-2

Zhang, W., Edwards, A., Flemington, E. K., and Zhang, K. (2013). Inferring polymorphism-induced regulatory gene networks active in human lymphocyte cell lines by weighted linear mixed model analysis of multiple RNA-Seq datasets. *PloS One* 8, e78868. doi: 10.1371/journal.pone.0078868

Zhang, X. F., Ou-Yang, L., Zhao, X. M., and Yan, H. (2016). Differential network analysis from cross-platform gene expression data. *Sci. Rep* 6, 34112. doi: 10.1038/srep34112

Zhang, Y., Ouyang, Z., and Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *Ann. App. Stat.* 11, 161–18. doi: 10.1214/16-AOAS998

Zhang, Z., Guo, X., and Yi, Y. (2015). Spectra of weighted scale-free networks. *Sci. Rep* 5, 17469. doi: 10.1038/srep17469

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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: 21 November 2019
doi: 10.3389/fgene.2019.01156

Reads Binning Improves Alignment-Free Metagenome Comparison

Kai Song[1†], Jie Ren[2†‡] and Fengzhu Sun[2]**

[1] School of Mathematics and Statistics, Qingdao University, Qingdao, China, [2] Quantitative and Computational Biology Program, University of Southern California, Los Angeles, CA, United States

Comparing metagenomic samples is a critical step in understanding the relationships among microbial communities. Recently, next-generation sequencing (NGS) technologies have produced a massive amount of short reads data for microbial communities from different environments. The assembly of these short reads can, however, be time-consuming and challenging. In addition, alignment-based methods for metagenome comparison are limited by incomplete genome and/or pathway databases. In contrast, alignment-free methods for metagenome comparison do not depend on the completeness of genome or pathway databases. Still, the existing alignment-free methods, d_2^S and d_2^*, which model *k*-tuple patterns using only one Markov chain for each sample, neglect the heterogeneity within metagenomic data wherein potentially thousands of types of microorganisms are sequenced. To address this imperfection in d_2^S and d_2^*, we organized NGS sequences into different reads bins and constructed several corresponding Markov models. Next, we modified the definition of our previous alignment-free methods, d_2^S and d_2^*, to make them more compatible with a scheme of analysis which uses the proposed reads bins. We then used two simulated and three real metagenomic datasets to test the effect of the *k*-tuple size and Markov orders of background sequences on the performance of these *de novo* alignment-free methods. For dependable comparison of metagenomic samples, our newly developed alignment-free methods with reads binning outperformed alignment-free methods without reads binning in detecting the relationship among microbial communities, including whether they form groups or change according to some environmental gradients.

Keywords: alignment-free methods, metagenomic samples, Markov model, reads binning, beta-diversity

OPEN ACCESS

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

Reviewed by:
Marc Sze,
Merck, United States
Bryan David Martin,
University of Washington, United States

****Correspondence:***
Kai Song
ksong@qdu.edu.cn
Fengzhu Sun
fsun@usc.edu

[†]These authors have contributed equally to this work

[‡]Present Address:
Jie Ren
Google Inc., Mountain View, CA, United States

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 January 2019*
Accepted: *22 October 2019*
Published: *21 November 2019*

Citation:
Song K, Ren J and Sun F (2019) Reads Binning Improves Alignment-Free Metagenome Comparison. Front. Genet. 10:1156. doi: 10.3389/fgene.2019.01156

INTRODUCTION

Understanding the impact of environmental factors on the composition of microbial communities, along with the effects of microbes on their hosts, is a crucial problem in microbiological studies. Traditional culture-dependent techniques can obtain pure isolates of individual microbes, but such techniques are low-throughput and can capture only a tiny fraction of microbes in a microbial community. With the rapid development of next-generation sequencing (NGS) technology, whole metagenome shotgun sequencing (WMGS) has become a widely used and powerful approach to investigate complex microbial communities (Qin et al., 2010; Qin et al., 2012; Xie et al., 2016; Mehta et al., 2018). Several large scale international metagenomics projects including the Human Microbiome Projects (HMP) (Lloyd-Price et al., 2019) and TARA ocean project (Brum et al., 2015; Sunagawa et al., 2015) have been carried out and most of the metagenomic samples have

metadata available. Metagenomic data provide the whole genetic information from microbial communities. A metagenomic sample usually contains millions of short reads, consisting of several hundred of base pairs, and each read is randomly sampled from a genomic region of a microbial genome in the community. Given the massive amount of metagenomic data, computational methods are in great demand to infer the relationships between microbes and environmental factors/hosts. Accurately quantifying the similarities and differences among microbial communities from multiple environments/hosts is one of the most important steps in metagenomic data analysis.

The general approach to analyze metagenomic data is based on alignment methods, such as the Smith-Waterman algorithm (Smith and Waterman, 1981) and BLAST (Altschul et al., 1990), both of which first map NGS reads to known genomes or pathways in existing public protein databases, such as non-redundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG), and then compare the abundance of different microbial organisms or functional categories between samples (Qin et al., 2010; Muegge et al., 2011; Qin et al., 2012). However, many microbial genomes and gene families are unknown, making it impossible to map all reads to the known genomes or pathways in many environments, in turn making the comparison of metagenomic samples incomplete, as suggested above. Based on the current literature, about 40% of unassigned reads, on average, exist in the human gut microbiome (Qin et al., 2010; Qin et al., 2012), and up to 50% of reads cannot be assigned to reference databases in ocean samples (Marchetti et al., 2012). Apart from alignment-based methods, assembly-based analytical methods reconstruct bacteria genomes by assembling short reads. However, assembly is time-consuming and challenging, especially for metagenomic samples because bacteria genomes can share similar regions, and a short read is not long enough to resolve the ambiguity. These limitations leave alignment-free methods as promising alternative approaches for microbial community comparison by eliminating the requirements of reference sequences or *de novo* assembly.

Although alignment-free methods can be defined as any methods that do not depend on sequence alignment, one of the major types of alignment-free methods is based on the frequencies of k-tuples (k-words or k-mers) as recently reviewed (Song et al., 2014; Zielezinski et al., 2017; Ren et al., 2018). A k-tuple is a segment consisting of consecutive nucleotide bases of length k. The effectiveness of these alignment-free methods for genome and metagenome comparison was based on the fact that relative k-tuple frequencies were similar across different regions of the same genome, but differed between genomes (Karlin et al., 1997). Similarly, the relative k-tuple frequencies for closely related genomes would be more similar than those between distantly related genomes. The alignment-free dissimilarity measures, d_2^S and d_2^*, were developed for high-throughput sequencing data comparison, and they were then used for phylogenetic tree construction (Song et al., 2013), followed by successful applications in the comparison of metagenomic samples (Jiang et al., 2012; Liao et al., 2016) and gene regulatory regions (Song et al., 2013), identification of horizontal gene transfer (Tang et al., 2018b) and virus-host interactions (Ahlgren et al., 2017), and improving contig binning for metagenomes (Wang et al., 2017). Recently, they have also been used to identify the geographic origin of white oak trees (Tang et al., 2018a) and sources of viruses (Li and Sun, 2018). A user-friendly interface for alignment-free genome and metagenome comparison, aCcelerated Alignment-FrEe (CAFÉ) (Lu et al., 2017b), has now been developed. Many other alignment-free methods have been developed including the delta-distance between dinucleotide relative frequencies of different genomes (Kariin and Burge, 1995; Karlin and Mrázek, 1997) and CVTree (Qi et al., 2004a; Qi et al., 2004b). Ren et al. (2018) and Zielezinski et al. (2017) presented the most recent reviews of alignment-free methods for genome and metagenome comparisons and their many applications (Zielezinski et al., 2017; Ren et al., 2018). Zielezinski et al. (2019) recently compared the performance of 74 alignment-free methods for protein sequence classification, gene tree inference, regulatory element detection, genome-based phylogenetic inference, and reconstruction of species trees under horizontal gene transfer, and recombination events. However, the authors did not evaluate their performance on metagenome comparison (Zielezinski et al., 2019).

While the previous alignment-free methods were successful in comparing metagenomic samples, these methods (Jiang et al., 2012; Liao et al., 2016) only considered metagenomics sequencing data as a whole from which to extract k-tuple frequencies and calculate their expectations using a common Markov model. However, microbial communities contain thousands of microorganisms and the relative abundance profiles of the microbial communities were shown to change across many environmental factors, such as geographic distance, temperature, oxygen, pH, and biotic factors (Lozupone and Knight, 2007; Steele et al., 2011; Philippot et al., 2013). Different microbial organisms have varied nucleotide frequencies; therefore, it is unreasonable to use only one Markov Chain to model the sequences in a microbial community and to calculate the probability of k-tuples. Instead, the present study posits that different Markov models can be used; accordingly, we first organized sequenced bacterial genomes and used them to construct the Markov models. These models were then used for grouping NGS reads into different bins, followed by extracting the k-tuples and calculating their expectation in each bin. Markov models have been used extensively for genome modeling (Narlikar et al., 2013), motif discovery (D'haeseleer, 2006), computational gene search (Lomsadze et al., 2005), classification of metagenomic sequences (Brady and Salzberg, 2009) and alignment-free sequence comparison (Chang and Wang, 2011). Next, we extended the definition of our previous alignment-free measures, d_2^S and d_2^*, to make them more compatible with a scheme of analysis that uses the proposed reads binning datasets. We then used two simulated and three real metagenomic datasets to test the effect of k-tuple size and Markov orders of background sequences on the performance of these *de novo* alignment-free methods. For dependable comparison of metagenomic samples, our alignment-free methods with reads binning outperformed alignment-free methods without reads binning in detecting the relationships among metagenomic samples whether they form groups or change according to environmental gradients. For

detecting group relationship among samples, the triplet distance between the inferred tree and the gold standard tree is reduced by over 10%. For detecting gradient relationship among the samples, the Pearson correlation coefficient (PCC) between the first principal coordinate and the gradient is increased by 10%. The software is available at https://github.com/songkai1987/MetaBin.

MATERIALS AND METHODS

The framework of our method is given in **Figure 1**. First, the bacterial sequences were divided into several bins and a Markov model is used to model the sequences in each bin. Second, each read in the metagenomics samples was assigned to the bin that has the highest probability of generating the sequence. Third, the *k*-tuple counts and their expectations were calculated in each bin of the NGS reads. The d_2^S and d_2^* (Eq. 1 and 2) were calculated between each pair of samples. Finally, the samples are clustered using the dissimilarity matrix obtained from d_2^S and d_2^* Details of each of the steps are given below.

The *k*-Tuple Count Vectors and Alignment-Free Comparison Measures

In our previous studies (Jiang et al., 2012; Song et al., 2013), the first step toward comparing metagenomic samples involved counting the number of occurrences of each *k*-tuple. Since a read could be from the forward or reverse strand of a genome, we considered each read together with its complement when calculating the occurrences of each *k*-tuple. Thus, for metagenomic data, we have a finite alphabet set *S={A,C,G,T}* and consider all possible *k*-tuples in the reads of metagenomic samples. Let $X=(X_1,X_2,...,X_{4^k})$ and $Y=(Y_1,Y_2,...,Y_{4^k})$ be the *k*-tuple count vectors of two metagenomic samples *X* and *Y*, respectively. Then, we define the centralized count variables by using Markov model-based expectation as

$$\overline{X}_i = X_i - n_X p_{X,i}$$

$$\overline{Y}_i = Y_i - n_Y p_{Y,i}$$

where n_X is the total count of *k*-tuples, and $p_{X,i}$ is the probability of *i*-th *k*-tuple under the Markov model of order *r*. The idea behind subtracting the expected *k*-tuple count from the observed count is that the *k*-tuples responsible for the similarity between microbial communities will stand out after subtraction. Then, the two measures d_2^S and d_2^* can be defined as

$$D_2^S(X,Y)=\sum_{i=1}^{4^k}\frac{\overline{X}_i\overline{Y}_i}{\sqrt{\overline{X}_i^2+\overline{Y}_i^2}}$$

$$d_2^S(X,Y)=\frac{1}{2}\left(1-\frac{D_2^S(X,Y)}{\sqrt{\sum_{i=1}^{4^k}\frac{\overline{X}_i^2}{\sqrt{\overline{X}_i^2+\overline{Y}_i^2}}\sum_{i=1}^{4^k}\frac{\overline{Y}_i^2}{\sqrt{\overline{X}_i^2+\overline{Y}_i^2}}}}\right) \quad (1)$$

and

$$D_2^*(X,Y)=\sum_{i=1}^{4^k}\frac{\overline{X}_i\overline{Y}_i}{\sqrt{n_X p_{X,i}}\sqrt{n_Y p_{Y,i}}}$$

$$d_2^*(X,Y)=\frac{1}{2}\left(1-\frac{D_2^*(X,Y)}{\sqrt{\sum_{i=1}^{4^k}\frac{\overline{X}_i^2}{n_X p_{X,i}}\sum_{i=1}^{4^k}\frac{\overline{Y}_i^2}{n_Y p_{Y,i}}}}\right) \quad (2)$$

The first statistic D_2^S is based on the observation by Shepp (Shepp, 2006) that for two independent normal random variables *X* and *Y* with mean zero, $XY/\sqrt{X^2+Y^2}$ is also normally distributed. The second statistic D_2^* is motivated by Pearson correlation where the mean and variance of each tuple are calculated based on Poisson distribution assumption

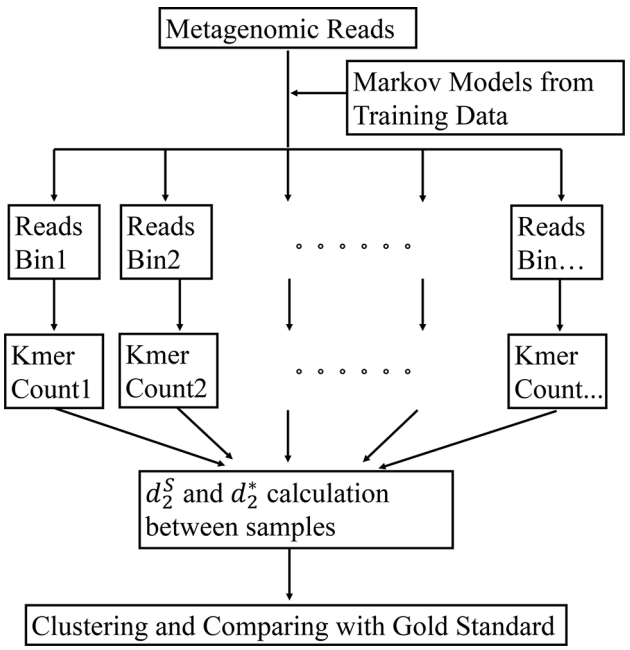


FIGURE 1 | The work flow of our approach. First, the Markov model for each bin is trained using the bacterial genomic sequences. Then, the metagenomic reads are binned to the group under which the sequence has the highest likelihood. The *k*-tuple counts and their expectations are calculated in each bin of the NGS reads. The d_2^S and d_2^* are calculated between each pair of samples. Finally, the samples are clustered using the dissimilarity matrix obtained from d_2^S and d_2^*.

for the *k*-tuples. When the two samples are more similar, the *k*-tuple frequency profiles are more similar and the values of D_2^S and D_2^* are higher. The ranges of D_2^S and D_2^* can depend on the nucleotide frequencies. In order to make their range independent of nucleotide frequencies, we normalize them to dissimilarities, d_2^S and d_2^*, respectively, so that they have a range between 0 and 1 according to the Cauchy inequality. When two samples are similar, the values of d_2^S and d_2^* are close to 0.

The Alignment-Free Measures Based on a Mixture of Markov Models Learned From Reads Bins

Metagenomic samples consist of a mixture of many different microbial genomes; thus, it is unreasonable to expect that all these reads can be modeled using only one single Markov model for each sample. To address this difficulty, we first group these reads into different bins. Then, we count the *k*-tuple vectors and obtain the expectation of each *k*-tuple for the reads in each bin individually.

We used the bacterial genomic sequences to train the Markov models. First, we calculated the guanine-cytosine (GC) frequency of each bacterial genomic sequence and then grouped these bacterial genomic sequences into different bins using the quantiles of the GC frequency distribution. Each bin has the same number of bacterial genomes. The Markov model for each bin was then constructed using the *k*-tuple vectors counted from all the genomic sequences in that bin. For a set of genomic sequences in a bin, let X_w be the count of *k*-tuple *w* of all these genomes and their complementary sequence. The Markov model of order *r* is defined as a $4^r \times 4$ matrix of transition probabilities. The transition probabilities can be estimated based on the *r*-tuples and $(r-1)$-tuples, and the estimated probability of observing nucleotide w_{r+1} given preceding nucleotides $w_1 w_2 \cdots w_r$ is $P_M(w_{r+1} \mid w_1 w_2 \cdots w_r) = \frac{X_{w_1 w_2 \cdots w_r w_{r+1}}}{X_{w_1 w_2 \cdots w_r}}$, where $X_{w_1 w_2 \cdots w_r}$ and $X_{w_1 w_2 \cdots w_r w_{r+1}}$ are the counts of *r*-tuple $w_1 w_2 \cdots w_r$ and $(r+1)$-tuple $w_1 w_2 \cdots w_r w_{r+1}$, respectively.

Once we have *C* different Markov models of order *r*, $(M_r^1, M_r^2, \cdots, M_r^C)$, to model the bacterial genomic sequences, we classify the reads in a metagenomic sample to the bins with the highest log-likelihood scores. In particular, suppose $Y = y_1 y_2 \cdots y_N$ represents a read of length N in a metagenomic sample; then, the log-likelihood of the read under the Markov chain M_r could be calculated as

$$LL(Y \mid M_r) = \sum_{i=1}^{N-r} \log P_{M_r}(y_{i+r} \mid y_i y_{i+1} \cdots y_{i+r-1})$$

Then, the classification of read could be defined as the model having the largest probability, or

$$l = \arg\max_{c=1,L,C} LL(Y \mid M_r^c) \tag{3}$$

where λ is the predicted bin to which the read belongs.

Next, we calculate the *k*-tuple count and its expectation in each bin of NGS reads. The centralized count variables by using Markov model-based expectation such that all *C* bins are combined are as follows:and

$$\overline{X}_w = \sum_{c=1}^{C} (X_w^c - n_X^c p_{X,w}^c) \tag{4}$$

$$\overline{Y}_w = \sum_{c=1}^{C} (Y_w^c - n_Y^c p_{Y,w}^c)$$

where *c* represents the calculation based on the *c*-th bin. Therefore, the two measures d_2^S and d_2^*, could be defined using the new version of $\overline{X}_w$ and $\overline{Y}_w$.

Comparison With Other Reads Binning Approaches Without Reference Genomes

In addition to the above reads binning method, we also considered creating reference-free reads binning by first assembling reads into contigs and grouping contigs into bins. Metagenomic reads are then classified to different bins based on their similarity to the contigs in those bins. MetaSPAdes (Bankevich et al., 2012; Nurk et al., 2017) was used to cross-assemble the reads in the simulated datasets using the default setting. Contig coverages [Fragments Per Kilobase per Million reads (FPKMs)] were determined by mapping reads with Bowtie2 (Langmead and Salzberg, 2012), using the default settings, and were averaged for each bin. Sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge (COCACOLA) (Lu et al., 2017a) and MetaBAT (Kang et al., 2015) were used to cluster these assembled contigs (≥500 bp) based on sequence tetra-nucleotide frequencies and contig coverages normalized by contig length and number of mapped reads in samples, respectively. MetaBAT performed better than other approaches in the CAMI study (Meyer et al., 2018). The simulated reads were mapped to the set of contigs using Burrows-Wheeler-Aligner (BWA) software (Li and Durbin, 2009) to obtain the classification labels. The unmapped reads were binned together as an extra bin. We calculated the *k*-tuple counts and their expectation in each bin and then calculated the values of d_2^S and d_2^*.

Comparison With Other Reads Binning Approaches With Reference Genomes

We compared our method with two reference genome-based reads binning approaches, Kraken (Wood and Salzberg, 2014) and MBMC (Wang et al., 2016), to classify the metagenomic reads. Kraken is a program for assigning taxonomic labels to metagenomic DNA sequences and it has been shown to perform better than other binning approaches, such as Megablast (Chen et al., 2015), PhymmBL (Brady and Salzberg, 2009), NBC (Rosen et al., 2008) and MetaPhlAn (Segata et al., 2012). The core of Kraken is a database consisting of *k*-tuples and the lowest common ancestor (LCA) of all organisms whose genomes

contain the *k*-tuples. Sequences are classified by querying the database for each *k*-tuple in a sequence, and then using the resulting set of LCA taxa to determine an appropriate label for the sequence. To compare with our method, the 100 bacterial genomes in simulations were used to construct the genome library for *k*-tuples and their LCAs in Kraken. MBMC is a recent approach for binning reads by measuring the similarity of reads to the trained Markov chains for different taxa using the ordinary least squares (OLS) method. Similarly, the 100 bacterial genomes in simulations were also used for constructing the Markov chains, respectively. Each of the two approaches was then used to classify reads into different bins individually. We calculated the *k*-tuple counts and their expectations in each bin to then calculate the values of d_2^S and d_2^*.

Beta-Diversity Analysis and Evaluation Methods

Detection of group relationships among metagenomic samples and the identification of external gradients driving shifts in microbial community structure are two major types of analytical tasks in microbial community comparison. Therefore, we evaluated the performance of our new alignment-free measures in metagenomic sample comparison by assessing how well they would detect the known group relationships or identify known environmental gradients.

For clustering analysis, we used the unweighted pair-group method with arithmetic means (UPGMA) algorithm (Murtagh, 1984) to cluster metagenomic samples based on the pairwise dissimilarity defined using our alignment-free measures, and then we compared the clustering tree with the true group relationship among the samples. We used the R package "phangorn" (Schliep, 2011) for clustering samples given the input of the pairwise dissimilarity matrix. The triplet distance was used to measure the distance between the tree built using our methods and the ground truth. Triplet distance was proposed by (Critchlow et al., 1996) as a measure for the distance between two rooted bifurcating phylogenetic trees, and it can be used for measuring the distance between binary (Critchlow et al., 1996) or non-binary trees (Bansal et al., 2011). This measure first decomposes the topologies of the input trees into triplets, i.e., all three-element subsets of the set of leaves, and then computes how many triplets of the two trees have different topologies. Because triplets are the basic building blocks of rooted and unrooted trees, in the sense that they are the smallest topological units that completely identify a phylogenetic tree, triplet-based distances provide a robust and fine-grained measure of the dissimilarities between trees (Bansal et al., 2011). This was finally developed into the TreeCmp toolbox (Bogdanowicz et al., 2012).

For the study of gradient relationships among the samples, the shift of metagenomic samples is visualized by PCoA (Principal Coordinates Analysis), which is a multidimensional scaling (MDS) method that converts between-sample dissimilarity matrix into two-dimensional, or three-dimensional, ordinates of samples and arranges the samples in ordinate space. We used the MASS package in R for PCoA (Anderson, 2003). Then, the influence of environmental gradient(s) on microbial communities could be investigated by calculating correlation, such as PCC, between the first principal coordinate and the gradient axis. In this way, the performance of the alignment-free methods could be evaluated, as long as the gradient driving microbial communities is known.

Simulated Metagenomic Datasets

We simulated two NGS metagenomic datasets using Next-generation Sequencing Simulator for Metagenomics (NeSSM) (Jia et al., 2013), which supports single-end and paired-end sequencing for both 454 and Illumina platforms, with paired-end short reads of length 150 bp in an Illumina MiSeq setting mode based on abundance profiles. Since 1) the database for reference genome is not complete and 2) new genomes can be discovered in the future, we mimic the situation by splitting the reference genomes by May 2015 such that the genomes before this date were used for training the Markov chain models, and the genomes after this date were used to simulate the metagenomic datasets for testing. A set of 100 bacterial species randomly sampled from the 5,865 sequenced bacterial reference genomes from NCBI was used for simulation (**Table S1**). We designed two sets of metagenomic samples representing the two types of relationships among samples as has been done in (Jiang et al., 2012): the group relationship involving species abundance levels of the samples belonging to different groups and the gradient relationship involving species abundance levels that change continuously with some environmental variables, such as temperature or location.

In Simulation 1, we simulated 60 samples belonging to three groups. For each group, we randomly chose 100 genomes and assigned the i-th genome with relative abundance generated from the power-law (Zipf's) distribution as $f(m;\alpha,N)=\frac{1/m^{\alpha}}{\sum_{n=1}^{N}1/n^{\alpha}}$, m = 1, 2, …, N, where N = 100, and α is the value of the exponent characterizing the distribution. We set α=0.3and generated three relative abundance vectors from power-law distribution by randomly ordering the 100 genomes as the centers of the three groups. We next added to each component the absolute value of a Gaussian noise with mean zero and variance equal to 10 times each component and then renormalized each component to sum to 1. Each relative abundance vector was randomized and renormalized 20 times, and a total of 60 relative abundance vectors were obtained. Then, we used the relative abundance vectors to simulate 60 metagenomic samples.

In **Simulation 2**, we generated 20 samples consisting of the same 100 genomes, and the relative abundance vector of 100 genomes was generated by the power law (Zipf's law) distribution as defined in the above simulation. In order to mimic the gradient model, the relative abundance vector shifts along a gradient axis of $\alpha$from 0.30 to 0.70 by step 0.02. Again, absolute values of Gaussian noises were added to each component of the 20 abundance vectors with mean 0 and standard deviation equal to the value of that component. The vectors were renormalized after adding the noises. We generated 20 metagenomic samples according to these relative abundance vectors using NeSSM.

In all simulations, we generated datasets at two sequencing depths: 0.1M and 0.5M sequencing reads per sample. At each setting, we generated 30 duplicated datasets to simulate possible stochastic effects in real NGS data.

Real Metagenomic Datasets

We analyzed three real shotgun metagenomic sequencing datasets published in recent years. For real datasets, we used all genomic sequences to train the Markov models.

The Human Gut Datasets

The first dataset includes 107 fecal microbiome samples from Asia (Kurokawa et al., 2007; Qin et al., 2012), Europe (Qin et al., 2010) and North America (Turnbaugh et al., 2009). The dataset includes samples from two countries (China and Japan, n = 45 and 13) in Asia, two countries (Denmark and Spain, n = 21 and 10) in Europe, and one country (USA, n = 18) in North America. The accession numbers for the samples are given in **Table S2** in the supplementary material. We investigated this dataset at two levels. First, we considered the samples from different continents and studied the relationships among these samples. Then, we considered the samples from different countries and studied the relationships among these samples with respect to their countries of origin.

The Human Microbiome Datasets

The second dataset includes 60 microbiome samples from four body sites: buccal mucosa, supragingival plaque, tongue dorsum and stool (Lloyd-Price et al., 2017). The accession numbers for the samples are given in **Table S3** in the supplementary material. We investigated the relationships among these microbial samples from different body sites.

The Soil Metagenomic Dataset

This dataset includes 16 soil metagenomic samples from 16 sites: 3 from hot deserts, 6 from Antarctic cold deserts, and 7 from temperate and tropical forests, a prairie grassland, a tundra, and a boreal forest (Fierer et al., 2012). The accession numbers of these samples are given in **Table S4** in the supplementary material. The sites span a wide range of ecologically distinct microbiomes to examine how cold desert soils compare with those from hot deserts, forests, prairie, and tundra. We investigated the relationships among these different ecologically distinct microbiomes and explored their relationship to environmental factors, such as pH values.

RESULTS

We conducted a series of computational experiments including both intensive simulations and real dataset analyses to study the effect of *k*-tuple-based alignment-free methods with or without reads binning on identifying group and gradient relationships of metagenomic samples. To accomplish this, we first simulated two types of metagenomic datasets to investigate the performance of our newly developed alignment-free measures d_2^S and d_2^*, and the effect of several factors, such as the *k*-tuple size and Markov orders of background sequences, on their performance. The simulated datasets were generated based on sampling reads from one hundred bacterial genomes randomly chosen from those detected after June 2015 with different abundance levels. The genomes discovered before May 2015 were used for training the Markov models for reads binning. We binned bacterial genomes by their GC content, and then, for each bin, we trained a Markov chain to model sequences in that bin. For reads in the simulated metagenomic samples, we classified them into different bins based on their likelihood evaluated under the corresponding Markov models [Eq. (3)]. The *k*-tuple frequency vectors were counted and normalized individually for each group [Eq. (4)]. Finally, the pairwise alignment-free dissimilarities, d_2^S and d_2^*, were computed between samples based on Eq. (1, 2), and β-diversity analysis was implemented to evaluate how well the true underlying relationship among samples could be recovered by our method. We also compared our newly developed methods with the original version of the alignment-free measures in (Jiang et al., 2012; Song et al., 2013) which were based on *k*-tuples, but without reads binning. In addition, we also compared our approach with two reference-free binning methods, COCACOLA and MetaBAT, and two other reference-based binning methods, Kraken and MBMC.

Simulation 1: Detecting Group Relationships Among Metagenomic Samples

In some situations, metagenomic samples may form different groups. For example, gut samples may group based on diet, and soil samples may group based on locations. In order to evaluate the ability of dissimilarity measures to detect such group relationships, we simulated datasets of 60 metagenomic samples belonging to three different groups (20 samples in each group) similar to the simulation design of (Jiang et al., 2012). Each sample was generated by simulating NGS reads from a mixture of 100 bacterial genomes detected after June 2015 with different abundance levels (see Materials and Methods for details).

We applied our newly developed alignment-free measures d_2^S and d_2^* to detect group relationships of the 60 samples by clustering analysis. We studied various factors, including the number of bins, the order of the Markov model for the background sequences, the tuple size *k*, and sequencing depth, all affecting the performance of d_2^S and d_2^* in recovering the group relationships among the samples. **Figure 2** showed that both d_2^S and d_2^* dissimilarity measures with reads binning outperform the original versions without reads binning. The best clustering result with the smallest triplet distance is obtained by d_2^S with reads binning using tuple size $k = 5$, Markov order 3 (**Figure 3**). To test if the lowest triplet distance is statistically significantly lower than the second lowest triplet distance, we generated 10 duplicated datasets to simulate possible stochastic effects in real NGS data and obtained the triplet distances between the inferred clustering and the reference cluster for each duplication. Using paired t-test, the resulting one side p-value is less than 0.0005 indicating that the lowest and the second lowest triplet distances are statistically significantly different. In **Table 1**, we fixed the tuple size at 5 for d_2^S and d_2^*, and compared the effect of reads binning number on recovering group relationships. The results showed that alignment-free methods without reads binning had

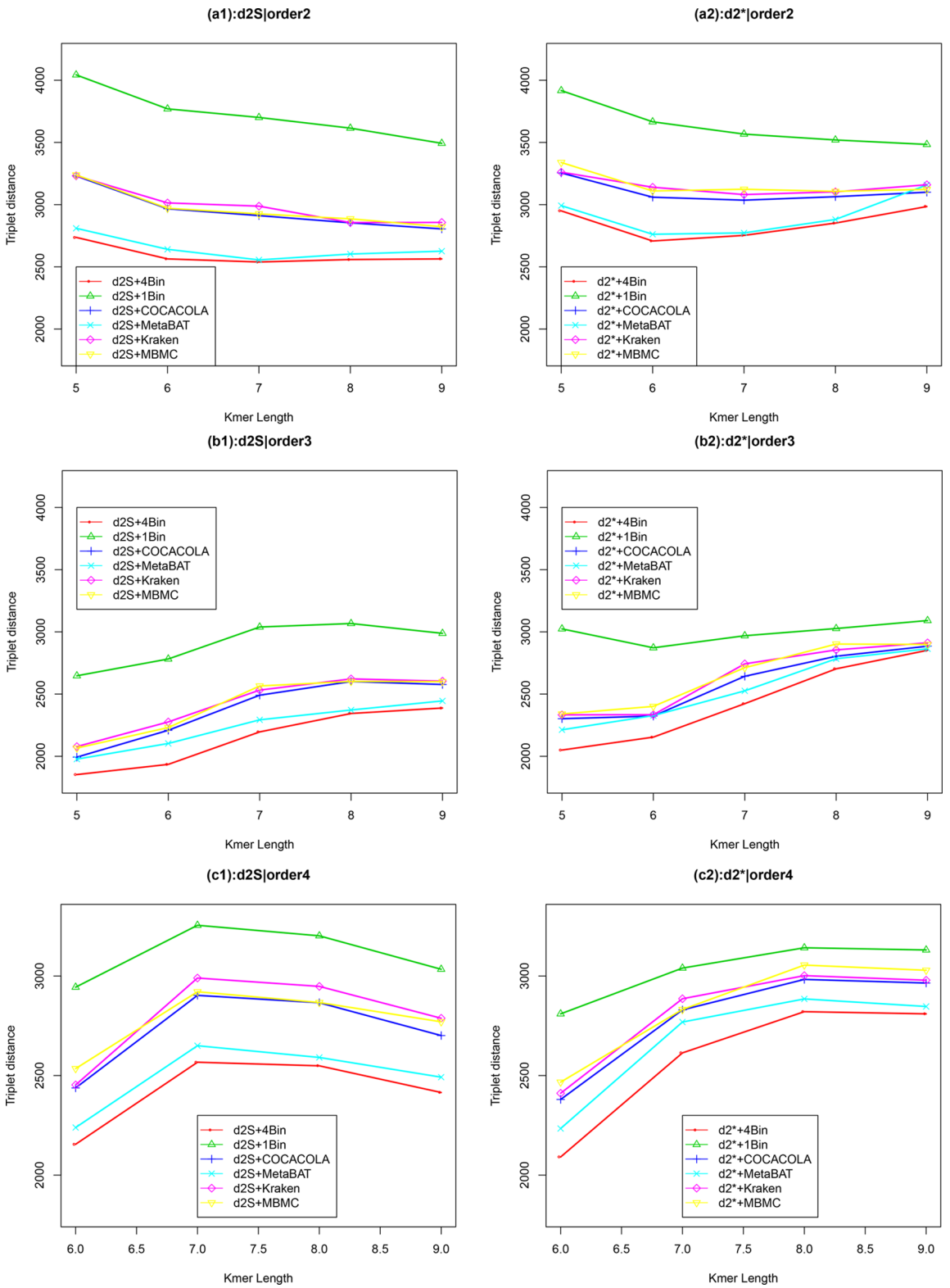


FIGURE 2 | The relative performance (triplet distance) of various reads binning methods in recovering group relationships of the metagenomic samples for Simulation 1 at sequencing depth of 500,000 NGS paired-end reads. The background sequence Markov orders were two (a1, a2), three (b1, b2), and four (c1, c2). The dissimilarity measures d_2^S and d_2^* with binning into 4 bins outperform other binning methods in most situations. The corresponding figures based on Markov order zero and one are presented as **Figure S2** in **Supplementary Material**.

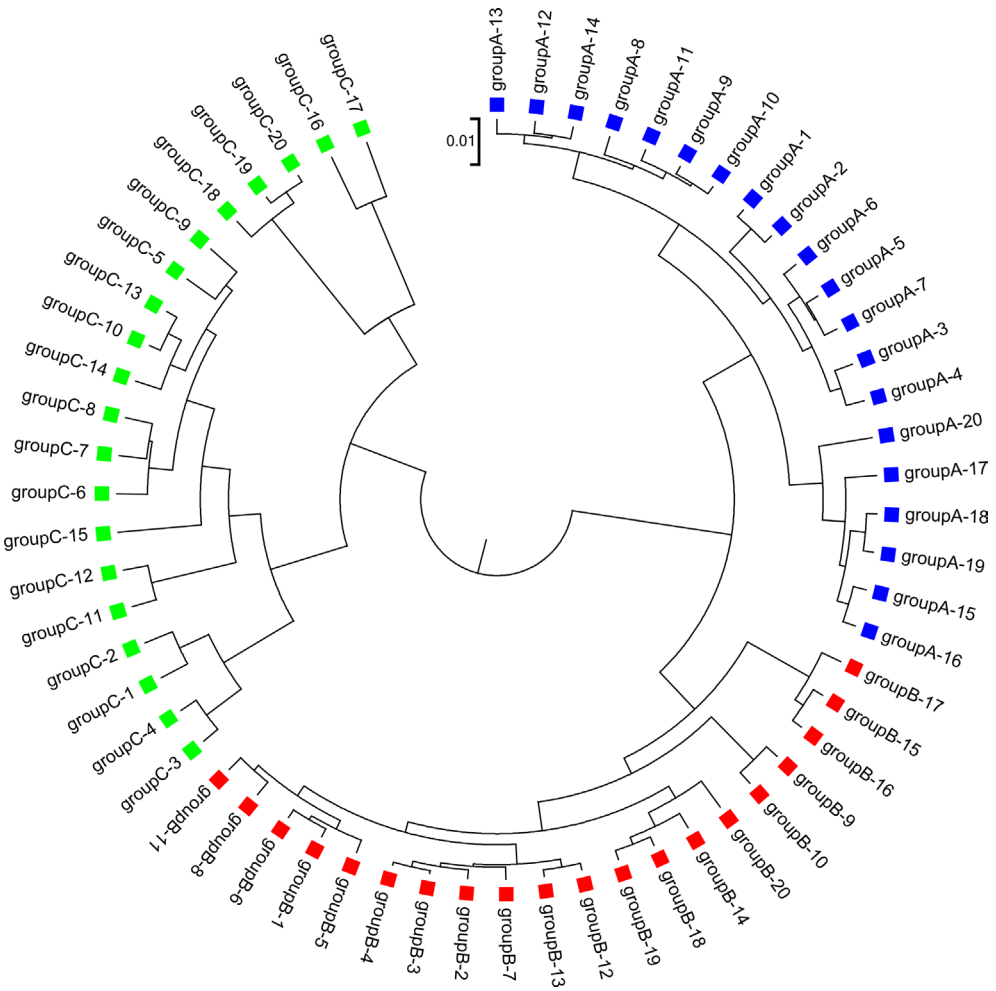


FIGURE 3 | The best clustering tree for the 60 simulated metagenomic samples in Simulation 1 based on the newly developed dissimilarity measure d_2^S with reads grouped to 4 bins, tuple size k = 5, and background sequence Markov order = 3.

TABLE 1 | The triplet distances between the reference and the clustering trees using various numbers of bins for the reads with tuple size k = 5 and background sequence Markov order from 0 to 3 for Simulation 1 at sequencing depth of 500,000 next-generation sequencing paired-end reads.

		No binning	2 bins	3 bins	4 bins	5 bins
d_2^S	order 0	3,535	2,634	2,635	2,634	2,633
	order 1	4,123	3,472	3,593	3,619	3,666
	order 2	4,043	2,867	2,846	2,737	2,726
	order 3	2,647	**1,852**	1,856	**1,853**	1,875
d_2^*	order 0	3,723	2,629	2,668	2,676	2,663
	order 1	4,183	3,833	3,977	3,992	4,042
	order 2	3,893	2,987	2,971	2,950	2,943
	order 3	2,986	2,087	2,020	2,050	2,045

The two lowest triplet scores are in boldface.

the largest values of triplet distance, i.e., the worst performance, compared to alignment-free methods with reads binning from 2 to 5 bins, which improved performance. Reads binning from 3, 4, or 5 bins could achieve similar performance. The simulations using a relatively shallow sequencing with 100,000 paired-end reads also gave results similar to those of deeper sequencing with 500,000 paired-end reads (**Figure S3**).

We next investigated the effects of sequencing errors on the performance of our methods and the results are shown in **Figure S1(a, b)** in the supplementary material. As expected, the sequencing errors could affect the accuracy of the reads assembly and contig binning, which in turn affect the clustering results. The triplet distance did not increase with sequencing error rate significantly until the sequencing error rate equals to 0.05 (**Figure S1**, p-value < 0.05 for t-tests). For reference, the sequencing error rates of Illumina and 454 platforms are ~0.001 or 0.01, respectively (Glenn, 2011), so sequencing errors only slightly impact the performance of the measures at the reported error rates for the NGS technologies.

We next considered other reference-independent and reference-dependent ways to construct Markov chain models. We cross-assembled the reads from the 60 metagenomic samples and used COCACOLA (Lu et al., 2017a) and MetaBAT (Kang et al., 2015), two reference-independent contig binning methods, to

bin these contigs, respectively. We also used two reference-based reads binning methods, Kraken (Wood and Salzberg, 2014) and MBMC (Wang et al., 2016), based on bacterial genomes to group the metagenomic reads into different bins. Then, Markov chain models were constructed for each contig bin, and reads were then classified in the same way to each contig bin based on their likelihood under different Markov models. We compared these reads binning schemes with our approach. **Figure 2** show the corresponding results. It can be seen that all these reads binning schemes are better than the original version without any reads binning procedure, but they do not perform as well as the above scheme based on binning from Markov chains.

Simulation 2: Revealing Environmental Gradients From Metagenomic Samples

The second simulation experiment was designed to evaluate the effectiveness of the alignment-free methods for analyzing gradient variation of microbial communities. A set of 20 metagenomic samples was generated by simulating NGS reads from 100 bacterial species also used in the above simulations with varying abundance levels. We designed the proportion of the 100 genomes to vary from sample 1 to sample 20 in a way that would mimic gradient variation across the samples, and then, we evaluated the performance of the alignment-free methods in terms of revealing such gradient variations from the metagenomics data.

Dissimilarity matrices were calculated using the alignment-free methods with different *k*-tuple sizes and Markov orders of background sequences as above. PCoA (Anderson, 2003), an effective approach to display β-diversity among multiple samples, mapped the 20 samples to a two-dimensional space. Then, the PCC was calculated between the first principal coordinate (PC1) given by PCoA and the predetermined gradient axis built into the simulation model. PCC can be taken as an index of how well the alignment-free method reveals the gradient variation in samples (see *Materials and Methods* for details). A higher PCC indicates better performance of the dissimilarity measure in recovering the gradient among the microbial samples.

Similar to Simulation 1, we generated two sequencing depths of 100,000 and 500,000 paired-end reads per sample. **Figure 4** showed the average PCC of the different dissimilarity measures at different tuple sizes and Markov orders of background sequences. Similar to the results in Simulation 1, reads binning improved the results compared to no binning for both alignment-free measures, d_2^S and d_2^*. The PCC values increased with tuple size and Markov order. For a fixed bin number of reads and tuple size, the PCC values increased more than 0.10 from order 0 to order 4, indicating that higher order Markov chains could model the genomic sequences better. The performance of d_2^* is slightly better than that of d_2^S for gradient detection. The best result with the largest PCC value was obtained by d_2^* with reads binning using tuple size $k = 9$ and background Markov order 4. To test if the highest PCC is statistically significantly higher than the second highest PCC, we generated 10 duplicated datasets to simulate possible stochastic effects in real NGS data and obtained the PCC for each duplication. Using paired t-test, the resulting one-sided p-value is less than 0.0005. In **Table 2**, we fixed the tuple size as 9 for d_2^S and d_2^*, and compared the effect of number of read bins on recovering gradient relationships. Again, results showed that the alignment-free methods without reads binning had the lowest values of PCC, i.e., worst performance, while methods with reads binning into 2 to 5 bins improved performance. For a given order of Markov chain, the PCCs corresponding to binning reads to 3, 4, or 5 bins are similar, indicating that that the number of reads bins does not markedly affect the performance of our methods when the bin number is at least 3. The simulations using a relatively shallow sequencing with 100,000 paired-end reads also gave results similar to those of deeper sequencing with 500,000 paired-end reads (**Figures S4** and **S5**). **Figure S1(c, d)** showed that the PCC values only decreased significantly when the sequencing error was 0.05 suggesting that sequencing errors only slightly impact the performance of the measures. **Figure 4** shows that all these reads binning schemes are better than the original version without any reads binning, but they do not perform as well as the above scheme based on binning from Markov chains.

Detecting Group Relationships Among Human Gut Samples

We applied the alignment-free methods to analyze human gut metagenomic datasets from different countries. These datasets include 107 fecal microbiome samples from Asia (Kurokawa et al., 2007; Qin et al., 2012), Europe (Qin et al., 2010) and North America (Turnbaugh et al., 2009). Two countries (China and Japan, n = 45 and 13) are from Asia, two countries (Denmark and Spain, n = 21 and 10) are from Europe, and one country (USA, n = 18) is from North America. In the simulation results, we found that the triplet distance and PCC values of the alignment-free dissimilarity measures d_2^S and d_2^* could achieve the best performance when the NGS reads were classified to four bins. Consequently, in the real data analysis, we used all the bacterial genomic sequences both before May 2015 and after June 2015 to construct four different Markov Models to bin these NGS reads.

TABLE 2 | The Pearson correlation between the first principal coordinate and the simulated environmental gradient using different numbers of bins for the reads with tuple size $k = 9$ and Markov order from 0 to 4 for Simulation 2 at sequencing depth of 500,000 next-generation sequencing paired-end reads.

		No binning	2 bins	3 bins	4 bins	5 bins
d_2^S	order 0	0.721	0.782	0.791	0.787	0.787
	order 1	0.769	0.855	0.852	0.851	0.849
	order 2	0.746	0.860	0.863	0.864	0.861
	order 3	0.805	0.896	0.893	0.887	0.844
	order 4	0.840	0.899	**0.907**	**0.907**	0.906
d_2^*	order 0	0.617	0.766	0.760	0.757	0.755
	order 1	0.724	0.871	0.870	0.871	0.871
	order 2	0.738	0.887	0.880	0.880	0.880
	order 3	0.807	0.904	0.903	0.904	0.901
	order 4	0.845	0.903	**0.914**	0.913	**0.914**

The two highest Pearson correlations are in boldface.

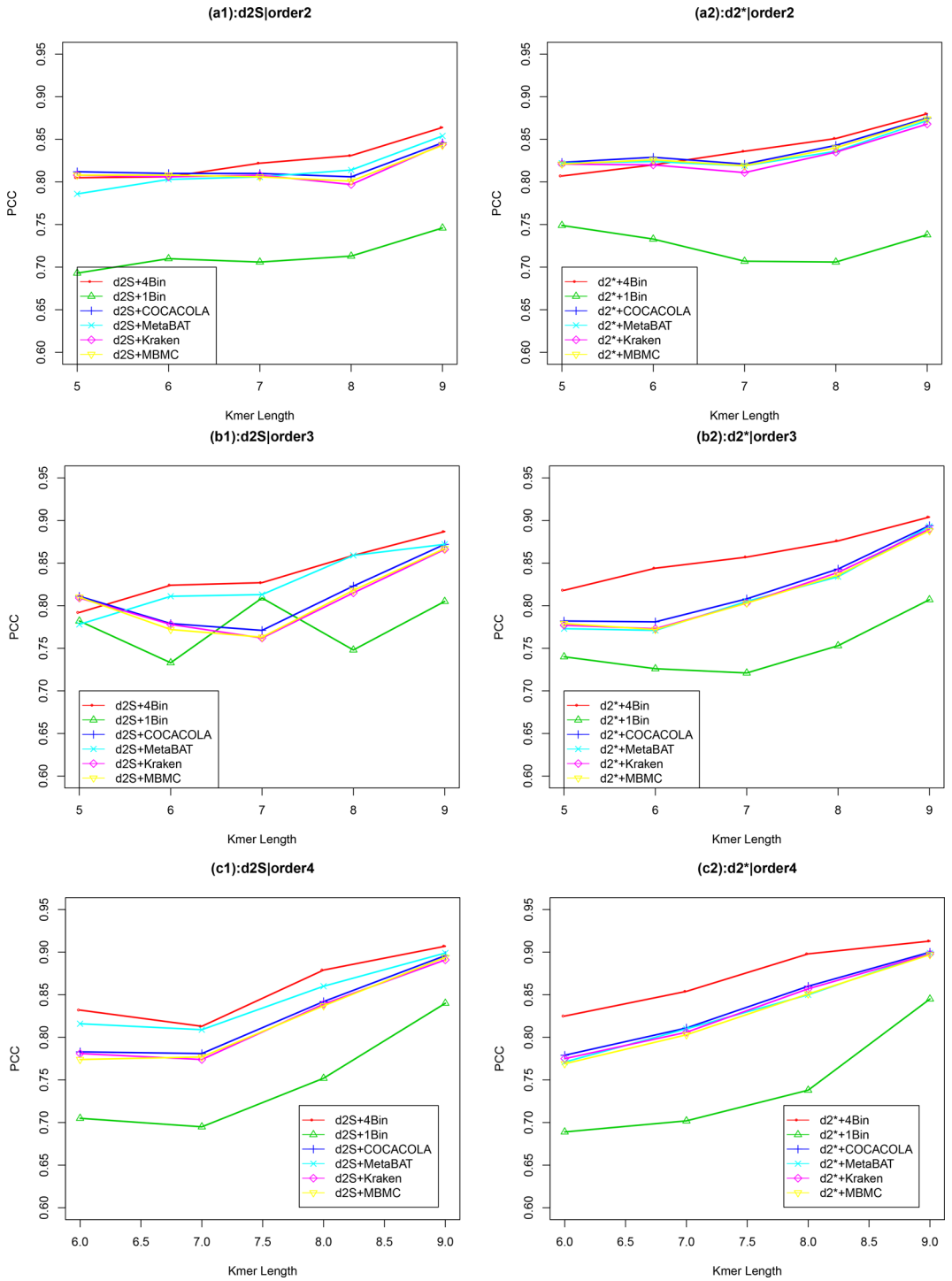


FIGURE 4 | The relative performance (Pearson correlation coefficient) of various reads binning methods in recovering gradient relationships of the metagenomic samples for Simulation 2 at sequencing depth of 500,000 next-generation sequencing paired-end reads. The background sequence Markov orders were two (a1, a2), three (b1, b2) and four (c1, c2). The dissimilarity measures d_2^S and d_2^* with binning into 4 bins outperform other binning methods in most situations. The corresponding figures based on Markov order zero and one are presented as **Figure S4** in **Supplementary Material**.

First, we used alignment-free measures, d_2^S and d_2^*, with tuple size 9 and Markov order 4 to explore the relationship among these human gut metagenomic samples. Similar to the simulation studies, we used UPGMA to cluster the samples based on the dissimilarity matrix, as defined by different dissimilarity measures based on sequence signatures. **Figure S6** showed that these human gut

samples could be clustered into four different groups labeled with different colors. The Japanese and American samples could be clearly separated from other groups with no overlaps. Most Chinese and European samples could be grouped separately, but with some overlaps. The samples from Denmark and Spain could not be distinguished from each other. A previous study (Costea et al., 2018) showed that the gut microbial community of both Chinese and European samples was enriched with *Firmicutes, Bacteroides and Prevotella*; however, the American samples all indicated a high-fat diet and were enriched with only *Bacteroides*. Therefore, both Chinese and European samples had similar microbial composition and should first be clustered together and then clustered again with the Japanese samples. The American samples have distinct gut microbial composition and should be separated from other samples.

We next calculated the triplet distance based on the four divided groups for d_2^S and d_2^*. The results of triplet distance scores for the different dissimilarity measures are summarized in **Table 3**. The smallest triplet distance score was achieved with d_2^S coupled with tuple size $k = 6$ and the fourth order Markov chain model of background sequences. When the order of Markov chains was four, the triplet distances were all lower than 30,000 for tuple size k from 6 to 9. In addition, triplet distance decreased with increasing Markov order for any fixed tuple size. The best performance was achieved when tuple size was $k = 6$ or 7 and Markov order = 4, similar to the k-tuple in Simulation 1. **Figure 5** showed the cluster tree using UPGMA for d_2^S with tuple size $k = 6$ and Markov order 4. **Table S5** showed the confusion matrix for d_2^S with tuple size $k = 6$ and Markov order 4. **Figure S7** showed the PCoA plot of these 107 samples. In this rooted tree, we found that American samples were separated from other samples and that the Japanese samples were separated from the Chinese and European samples. Although some European samples were mixed with the Chinese samples, most European samples clustered together.

Detecting Group Relationships Among Human Body Sites

We applied the alignment-free methods to analyze human metagenomic datasets from four body sites: buccal mucosa, supragingival plaque, tongue dorsum, and stool (Lloyd-Price et al., 2017). Each body site had fifteen samples. We calculated the pairwise d_2^S and d_2^* dissimilarities for any pair of samples

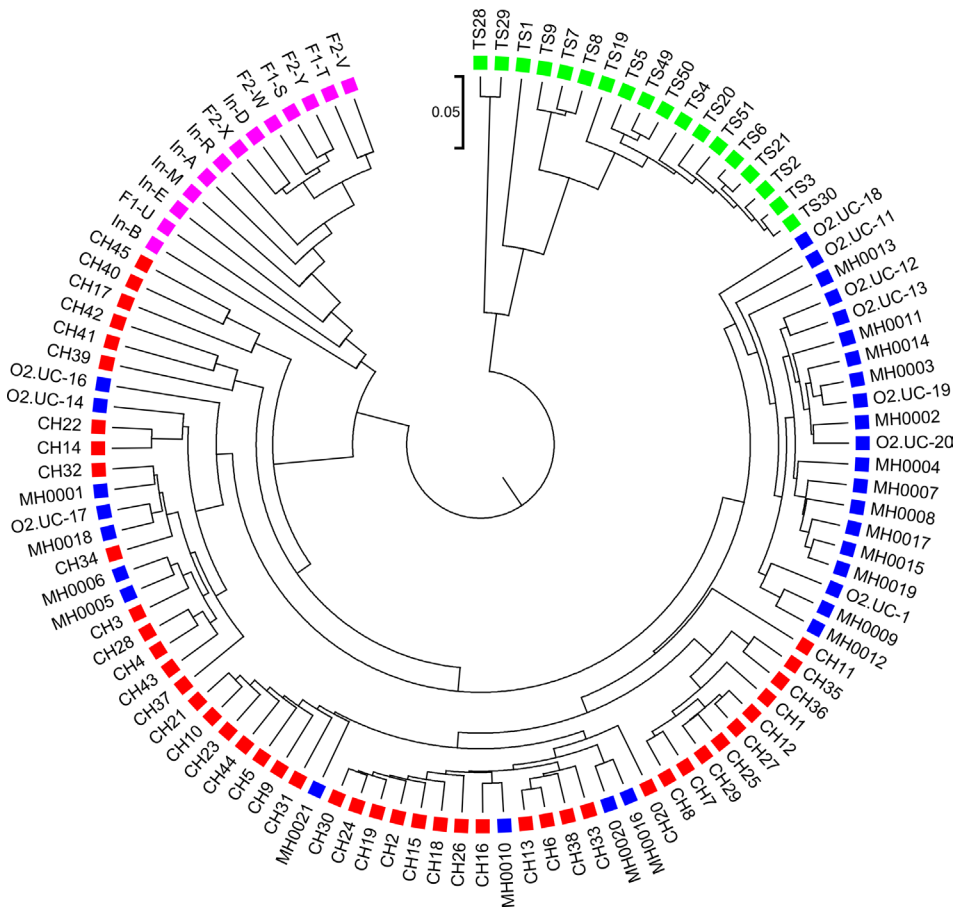


FIGURE 5 | The best clustering tree for the 107 human fecal metagenomic samples based on the newly developed dissimilarity measure d_2^S with tuple size $k = 6$ and background sequence Markov order = 4. Red squares: Chinese samples; blue squares: European samples; purple squares: Japanese samples; green squares: American samples.

TABLE 3 | The triplet distance between the reference and the clustering trees for the 107 human fecal metagenomic samples using various reads binning methods with tuple size k = 5–9 and background sequence Markov order from 0 to 4.

	k	5	6	7	8	9
d_2^S without reads binning	order 0	39,281	36,237	34,049	32,908	32,192
	order 1	38,129	35,070	33,306	32,455	32,149
	order 2	34,430	32,511	31,631	31,308	31,645
	order 3	32,124	31,154	31,629	31,738	32,162
	order 4	–	29,841	30,576	31,246	32,063
d_2^S with 4 bins	order 0	36,468	33,781	31,822	30,735	30,335
	order 1	35,568	32,215	30,569	30,114	30,287
	order 2	29,511	29,006	28,556	28,625	29,436
	order 3	31,112	30,130	29,350	29,468	30,256
	order 4	–	**26,890**	**26,962**	28,102	29,587
d_2^* without reads binning	order 0	49,732	46,565	42,415	37,998	34,036
	order 1	48,002	45,070	41,444	38,009	33,151
	order 2	43,132	40,134	38,055	33,539	32,171
	order 3	39,180	37,056	34,468	32,912	32,183
	order 4	–	34,656	33,829	33,215	33,054
d_2^* with 4 bins	order 0	46,942	44,312	40,504	36,556	32,285
	order 1	44,447	41,995	38,726	35,658	31,474
	order 2	37,515	35,859	33,896	30,249	30,154
	order 3	38,555	35,964	32,126	30,965	30,689
	order 4	–	31,816	30,064	30,031	30,799

The two lowest triplet distances are in boldface.

TABLE 4 | The triplet distance between the reference and the clustering trees for the 60 human metagenomic samples across four body sites using various reads binning methods with tuple size k = 5–9 and background sequence Markov order from 0 to 4.

	K	5	6	7	8	9
d_2^S without reads binning	order 0	4,536	4,153	3,696	3,306	2,986
	order 1	4,245	3,906	3,887	3,598	3,243
	order 2	3,945	3,657	3,257	3,010	2,798
	order 3	3,116	2,954	2,779	2,638	2,497
	order 4	–	2,215	2,275	2,315	2,382
d_2^S with 4 bins	order 0	4,342	3,982	4,407	4,073	3,672
	order 1	4,048	3,803	3,544	3,263	3,010
	order 2	3,843	3,541	3,248	3,061	2,868
	order 3	2,960	2,812	2,697	2,573	2,469
	order 4	–	**2,167**	**2,180**	2,206	2,261
d_2^* without reads binning	order 0	5,281	5,533	6,068	6,419	6,827
	order 1	4,534	5,244	6,069	6,610	6,841
	order 2	4,409	4,744	5,235	5,611	6,254
	order 3	3,800	4,286	5,034	5,861	6,387
	order 4	–	4,057	4,898	5,719	6,269
d_2^* with 4 bins	order 0	4,640	5,104	5,907	6,436	6,871
	order 1	4,527	5,034	5,837	6,178	6,658
	order 2	4,313	4,978	5,895	6,553	6,879
	order 3	3,496	4,080	4,907	5,836	6,396
	order 4	–	3,823	4,726	5,683	6,315

The two lowest triplet distances are in boldface

and build a hierarchical clustering tree. We next calculated the triplet distance between the clustering tree with the four divided groups based on body sites. **Table 4** showed that the smallest triplet distance score was achieved with d_2^S coupled with tuple size k = 6 and the fourth order Markov model of background sequences. **Figure 6** showed the cluster tree using UPGMA for d_2^S with tuple size k = 6 and Markov order 4. **Table S6** showed the confusion matrix for d_2^S with tuple size k = 6 and Markov order 4. In this rooted tree, we found that supragingival plaque and tongue dorsum samples were first grouped together and then clustered with the stool samples and buccal mucosa samples, consistent with the results from a previous study (Lloyd-Price et al., 2017).

Detecting Group and Gradient Variations in Soil Metagenomic Data

We next applied the alignment-free methods to analyze the metagenomic data of soil microbial communities collected from different geographic locations, spanning a wide range of ecologically distinct biomes, to examine how cold desert soils would compare with hot desert soils, forests, prairie, and tundra (Fierer et al., 2012).

The 16 soil samples form three ecologically distinct groups: hot deserts (n = 3), cold deserts (n = 6), and worldwide forests (n = 7). We conducted clustering analysis with sequence signatures of these samples and used triplet distance to study how well the grouping information was revealed (**Table 5**). Again, for all tuple size values, it can be seen that the performance of the alignment-free methods improved along with reads binning. Under reads binning, d_2^* coupled with tuple size k = 6 and the fourth order Markov model of background sequences achieved the best performance (**Tables 5** and **S7**, **Figure 7**). We observed that the three major groups identified by the alignment-free methods, d_2^S and d_2^*, reflected three major ecologically distinct conditions. The main factor that differentiates these soil samples is pH which, in polar and hot deserts, is higher than 7.00, but in worldwide forests lower than 7.00. These three groups of samples had different ranges of pH values. The pH of polar desert ranged from 8.15 to 9.95, while the pH values of hot desert ranged from 7.90 to 8.38. The pH values of worldwide forests ranged from 4.12 to 6.37. In the forest soil samples, the two samples from tropical forest (PE6) and Arctic tundra (TL1) with lowest pH values (4.12 and 4.58) were first clustered together and then clustered again with other forest samples. In order to test whether pH was the main environmental driver of microbial community composition, we tested the correlation between pH values and the first principal coordinate of these samples, and a highly significant negative correlation was found, as shown in **Figure S8** (Pearson correlation = −0.856, p-values = 0.0001). We also examined the correlation among the first to fourth principal coordinate of these samples with other environmental factors, including mean annual precipitation (MAP), mean

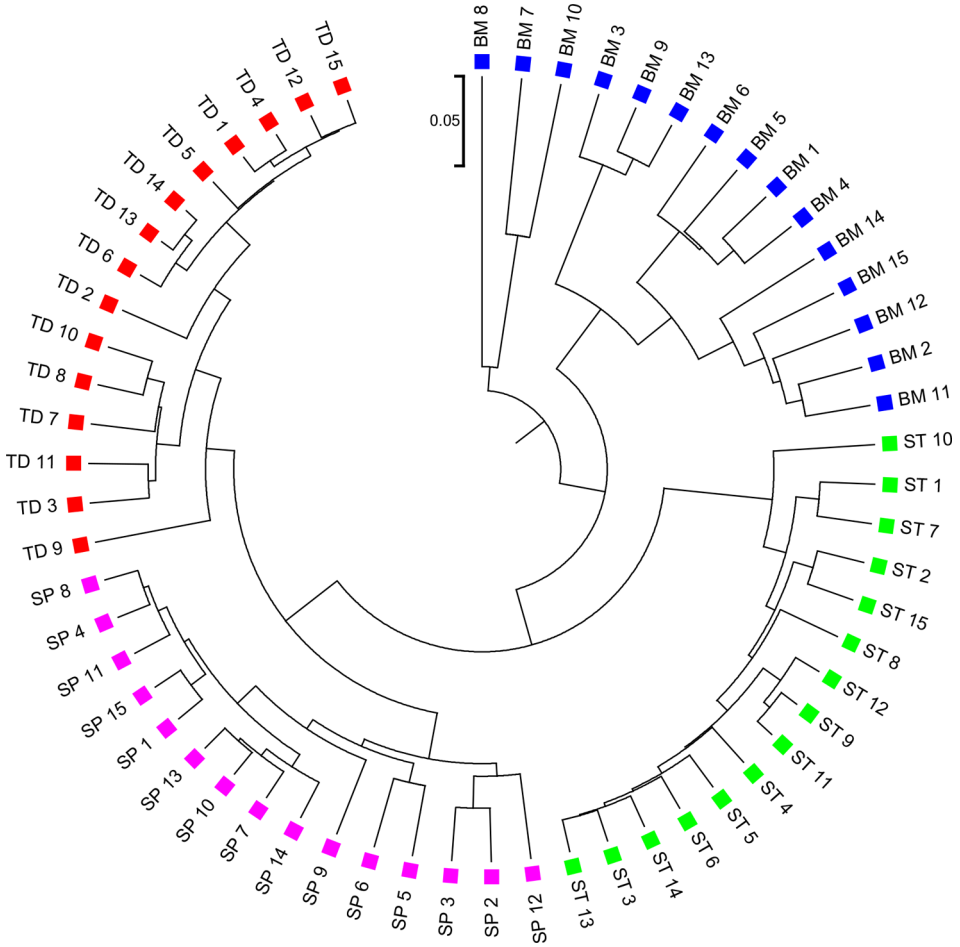


FIGURE 6 | The best clustering tree for the 60 human microbiome samples from four body sites based on newly developed dissimilarity measure d_2^S with tuple size k = 6 and background sequence Markov order = 4. Red squares: Tongue dorsum; Blue squares: Buccal mucosa; Purple squares: Supragingival plaque; Green squares: Stool.

annual temperature (MAT), organic Carbon content (%C), Nitrogen content (%N), and Carbon : Nitrogen ratio (C:N ratio). The first principal coordinate was also associated with the %C, %N, and C:N ratio (p-values < 0.01). But for the second, third, and fourth principal coordinates, the associations were not significant (**Table S8**).

DISCUSSION

In this study, we developed new alignment-free measures d_2^S and d_2^* for the comparison of metagenomes that model metagenomic reads as from a mixture of multiple Markov chains. We investigated the applications of the new alignment-free measures to compare metagenomic samples. Because of the high complexity of metagenomic data, the previous version of alignment-free measures d_2^S and d_2^* in (Jiang et al., 2012) that used only one background Markov model could not capture data heterogeneity. We proposed to first group reads in metagenomic samples into various bins using different Markov models. Then, k-tuple frequency vectors were counted and normalized individually in each bin. With the newly developed mixture model for computing the k-tuple expectations, we found that the modified d_2^S and d_2^* measures with reads binning outperformed the old ones in terms of recovering group and gradient relationships among samples from different environments. We extensively tested the methods on two sets of simulated metagenomic data and two sets of real metagenomic data, including metagenomes of human gut samples and worldwide soil samples. The effects of tuple size k, Markov order, and the bin number on the performance of our newly developed alignment-free measures were investigated, and the optimal ranges of those parameters were obtained.

There are several limitations of the current study. First, the performance of the new d_2^S and d_2^* measures depends on the number of bins for the reads. In this study, we let the number of bins be 1 to 5 and found that the optimal number of bins for the reads is between 3 and 5 in both simulation and real studies. In practice, we suggest setting the number of bins for the reads as 4. More studies are needed to see if

TABLE 5 | The triplet distance between the reference and the clustering trees for the 16 soil metagenomic samples from three ecologically distinct groups using various reads binning methods with tuple size k = 5–9 and background sequence Markov order from 0 to 4.

	k	5	6	7	8	9
d_2^S without reads binning	order0	127	121	117	115	115
	order1	110	111	112	113	110
	order2	113	118	116	115	115
	order3	114	113	119	120	123
	order4	–	117	117	118	124
d_2^S with 4 bins	order0	129	124	124	124	122
	order1	120	121	119	119	118
	order2	114	116	119	121	123
	order3	**108**	111	119	121	123
	order4	–	**108**	117	115	121
d_2^* without reads binning	order0	115	125	124	120	116
	order1	119	110	111	117	117
	order2	122	120	119	121	141
	order3	124	116	123	136	140
	order4	–	116	130	142	149
d_2^* with 4 bins	order0	129	126	124	122	116
	order1	122	119	117	119	135
	order2	121	120	120	129	144
	order3	112	112	121	142	143
	order4	–	119	135	145	153

The two lowest triplet scores are in boldface

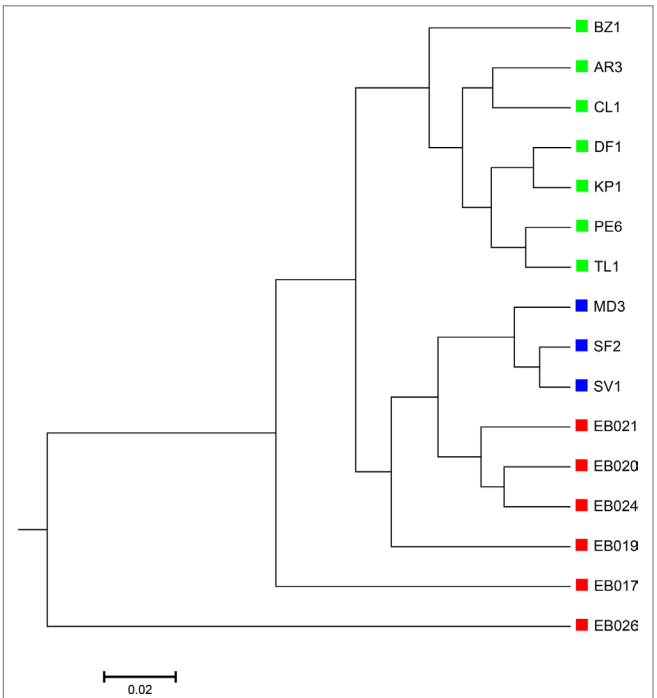


FIGURE 7 | The best clustering tree for the 16 soil metagenomic samples from three ecologically distinct groups based on the newly developed dissimilarity measure d_2^S coupled with tuple size k = 6 and background sequence Markov order = 4. Red squares: polar desert samples; blue squares: hot desert samples; green squares: forest samples.

this conclusion is robust for most comparative studies of metagenomic datasets. Second, the tuple size k can markedly impact the performance of the new d_2^S and d_2^* measures, and the optimal range of k can increase with sequencing depth. In general, the tuple size from 6 to 9 can give reasonable results. Third, the optimal range of Markov order is between 3 and 4 in most of our studies. Finally, d_2^S and d_2^* have similar performance, but d_2^S slightly outperforms d_2^* in most studied scenarios. This result is consistent with the finding that the old version of d_2^S slightly outperforms the old version of d_2^* without reads binning.

In this study, we focused on the comparison of metagenomic samples using alignment-free methods with reads binning. However, compared to alignment-based methods for mapping the reads to known genome or pathway databases and then comparing the genome and pathway abundance profiles, alignment-free methods cannot give insights about genomes and pathways responsible for the differences. From this perspective, we can say that alignment-free and alignment-based methods for metagenome comparison complement each other and should be used interactively to understand the dynamics of microbial communities.

AUTHOR CONTRIBUTIONS

KS and FS conceived of the project and developed the methods. KS and JR performed the computations. All authors discussed the results and contributed to the final manuscript.

FUNDING

The research was supported by the National Natural Science Foundation of China (11701546), U.S. National Institutes of Health (R01GM120624), and National Science Foundation (DMS-1518001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01156/full#supplementary-material

REFERENCES

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., and Sun, F. Z. (2017). Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45, 39–53. doi: 10.1093/nar/gkw1002

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Of Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anderson, M. (2003).*PCO: a FORTRAN computer program for principal coordinate analysis*. New Zealand: Department of Statistics, University of Auckland.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Of Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bansal, M. S., Dong, J. R., and Fernandez-Baca, D. (2011). Comparing and aggregating partially resolved trees. *Theor. Comput. Sci.* 412, 6634–6652. doi: 10.1016/j.tcs.2011.08.027

Bogdanowicz, D., Giaro, K., and Wrobel, B. (2012). TreeCmp: comparison of trees in polynomial time. *Evol. Bioinf.* 8, 475–487. doi: 10.4137/EBO.S9657

Brady, A., and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* 6, 673–676. doi: 10.1038/nmeth.1358

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., et al. (2015). Patterns and ecological drivers of ocean viral communities. *Sci.* 348, 1261498. doi: 10.1126/science.1261498

Chang, G. S., and Wang, T. M. (2011). Weighted relative entropy for alignment-free sequence comparison based on markov model. *J. Of Biomol. Struct. Dynamics* 28, 545–555. doi: 10.1080/07391102.2011.10508594

Chen, Y., Ye, W., Zhang, Y., and Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.* 43, 7762–7768. doi: 10.1093/nar/gkv784

Costea, P. I., Hildebrand, F., Arumugam, M., Backhed, F., Blaser, M. J., Bushman, F. D., et al. (2018). Enterotypes in the landscape of gut microbial community composition (vol 3, pg 8, 2017). *Nat. Microbiol.* 3, 388–388. doi: 10.1038/s41564-018-0114-x

Critchlow, D. E., Pearl, D. K., and Qian, C. L. (1996). The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.* 45, 323–334. doi: 10.1093/sysbio/45.3.323

D'haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nat. Biotechnol.* 24, 959–961. doi: 10.1038/nbt0806-959

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., et al. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Of Natl. Acad. Of Sci. Of U. States Of America* 109, 21390–21395. doi: 10.1073/pnas.1215210110

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

Jia, B., Xuan, L. M., Cai, K. Y., Hu, Z. Q., Ma, L. X., and Wei, C. C. (2013). NeSSM: a next-generation sequencing simulator for metagenomics. *PloS One* 8, e75448. doi: 10.1371/journal.pone.0075448

Jiang, B., Song, K., Ren, J., Deng, M. H., Sun, F. Z., and Zhang, X. G. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 13, 730. doi: 10.1186/1471-2164-13-730

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165. doi: 10.7717/peerj.1165

Kariin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends In Genet.* 11, 283–290. doi: 10.1016/S0168-9525(00)89076-9

Karlin, S., and Mrázek, J. (1997). Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci.* 94, 10227–10232. doi: 10.1073/pnas.94.19.10227

Karlin, S., Mrazek, J., and Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Of Bacteriol.* 179, 3899–3913. doi: 10.1128/jb.179.12.3899-3913.1997

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., et al. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 14, 169–181. doi: 10.1093/dnares/dsm018

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–U354. doi: 10.1038/nmeth.1923

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

Li, H., and Sun, F. Z. (2018). Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. *Sci. Rep.* 8, 10032. doi: 10.1038/s41598-018-28308-x

Liao, W., Ren, J., Wang, K., Wang, S., Zeng, F., Wang, Y., et al. (2016). Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length markov chains. *Sci. Rep.* 6, 37243. doi: 10.1038/srep37243

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. *Nat.* 550, 61. doi: 10.1038/nature23889

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. *Nat.* 569, 655. doi: 10.1038/s41586-019-1237-9

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 33, 6494–6506. doi: 10.1093/nar/gki937

Lozupone, C. A., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc. Of Natl. Acad. Of Sci. Of U. States Of America* 104, 11436–11440. doi: 10.1073/pnas.0611525104

Lu, Y. Y., Chen, T., Fuhrman, J. A., and Sun, F. Z. (2017a). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinf.* 33, 791–798. doi: 10.1093/bioinformatics/btw290

Lu, Y. Y., Tang, K. J., Ren, J., Fuhrman, J. A., Waterman, M. S., and Sun, F. Z. (2017b). CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Res.* 45, W554–W559. doi: 10.1093/nar/gkx351

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., et al. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc. Of Natl. Acad. Of Sci. Of U. States Of America* 109, E317–E325. doi: 10.1073/pnas.1118408109

Mehta, R. S., Abu-Ali, G. S., Drew, D. A., Lloyd-Price, J., Subramanian, A., Lochhead, P., et al. (2018). Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3, 347–355. doi: 10.1038/s41564-017-0096-0

Meyer, F., Hofmann, P., Belmann, P., Garrido-Oter, R., Fritz, A., Sczyrba, A., et al. (2018). AMBER: assessment of metagenome binners. *GigaScience* 7, giy069. doi: 10.1093/gigascience/giy069

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., Gonzalez, A., Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Sci.* 332, 970–974. doi: 10.1126/science.1198719

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: State of the art. *Comput. Stat Q.* 1, 101–113.

Narlikar, L., Mehta, N., Galande, S., and Arjunwadkar, M. (2013). One size does not fit all: on how markov model order dictates performance of genomic sequence analyses. *Nucleic Acids Res.* 41, 1416–1424. doi: 10.1093/nar/gks1285

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and Van Der Putten, W. H. (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* 11, 789–799. doi: 10.1038/nrmicro3109

Qi, J., Luo, H., and Hao, B. (2004a). CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.* 32, W45–W47. doi: 10.1093/nar/gkh362

Qi, J., Wang, B., and Hao, B.-I. (2004b). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* 58, 1–11. doi: 10.1007/s00239-003-2493-7

Qin, J. J., Li, R. Q., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nat.* 464, 59–U70. doi: 10.1038/nature08821

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nat.* 490, 55–60. doi: 10.1038/nature11450

Ren, J., Bai, X., Lu, Y. Y., Tang, K., Wang, Y., Reinert, G., et al. (2018). Alignment-free sequence analysis and applications. *Annu. Rev. Biomed. Data Sci.* 1, 93–114. doi: 10.1146/annurev-biodatasci-080917-013431

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using N-Mer frequency profiles. *Adv. In Bioinf.* 2008. doi: 10.1155/2008/205969

Schliep, K. P. (2011). Phangorn: phylogenetic analysis in R. *Bioinf.* 27, 592–593. doi: 10.1093/bioinformatics/btq706

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. doi: 10.1038/nmeth.2066

Shepp, L. (2006). Normal functions of normal random variables. *Siam Rev.* 6, 459–460. doi: 10.1137/1006100

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Of Mol. Biol.* 147, 195–197. doi: 10.1016/0022-2836(81)90087-5

Song, K., Ren, J., Zhai, Z. Y., Liu, X. M., Deng, M. H., and Sun, F. Z. (2013). Alignment-free sequence comparison based on next-generation sequencing reads. *J. Of Comput. Biol.* 20, 64–79. doi: 10.1089/cmb.2012.0228

Song, K., Ren, J., Reinert, G., Deng, M. H., Waterman, M. S., and Sun, F. Z. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings In Bioinf.* 15, 343–353. doi: 10.1093/bib/bbt067

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* 5, 1414–1425. doi: 10.1038/ismej.2011.24

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Sci.* 348, 1261359. doi: 10.1126/science.1261359

Tang, K., Ren, J., Cronn, R., Erickson, D. L., Milligan, B. G., Parker-Forney, M., et al. (2018a). Alignment-free genome comparison enables accurate geographic sourcing of white oak DNA. *BMC Genomics* 19, 896. doi: 10.1186/s12864-018-5253-1

Tang, K. J., Lu, Y. Y., and Sun, F. Z. (2018b). Background adjusted alignment-free dissimilarity measures improve the detection of horizontal gene transfer. *Front. In Microbiol.* 9, 711. doi: 10.3389/fmicb.2018.00711

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nat.* 457, 480–U487. doi: 10.1038/nature07540

Wang, Y., Hu, H., and Li, X. (2016). MBMC: an effective Markov chain approach for binning metagenomic reads from environmental shotgun sequencing projects. *Omics: A J. Integr. Biol.* 20, 470–479. doi: 10.1089/omi.2016.0081

Wang, Y., Wang, K., Lu, Y. Y., and Sun, F. Z. (2017). Improving contig binning of metagenomic data using d(2)(S) oligonucleotide frequency dissimilarity. *BMC Bioinf.* 18, 425. doi: 10.1186/s12859-017-1835-1

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

Xie, H. L., Guo, R. J., Zhong, H. Z., Feng, Q., Lan, Z., Qin, B. C., et al. (2016). Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* 3, 572–57+. doi: 10.1016/j.cels.2016.10.004

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* 18, 186. doi: 10.1186/s13059-017-1319-7

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., et al. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* 20, 144. doi: 10.1186/s13059-019-1755-7

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: 13 March 2020
doi: 10.3389/fgene.2020.00148

An Information-Based Approach for Mediation Analysis on High-Dimensional Metagenomic Data

Kyle M. Carter[1], *Meng Lu*[1], *Hongmei Jiang*[2] *and Lingling An*[1,3,4]*

[1] Interdiciplanary Program in Statistics and Data Science, The University of Arizona, Tucson, AZ, United States, [2] Department of Statistics, Northwestern University, Evanston, IL, 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

The human microbiome plays a critical role in the development of gut-related illnesses such as inflammatory bowel disease and clinical pouchitis. A mediation model can be used to describe the interaction between host gene expression, the gut microbiome, and clinical/health situation (e.g., diseased or not, inflammation level) and may provide insights into underlying disease mechanisms. Current mediation regression methodology cannot adequately model high-dimensional exposures and mediators or mixed data types. Additionally, regression based mediation models require some assumptions for the model parameters, and the relationships are usually assumed to be linear and additive. With the microbiome being the mediators, these assumptions are violated. We propose two novel nonparametric procedures utilizing information theory to detect significant mediation effects with high-dimensional exposures and mediators and varying data types while avoiding standard regression assumptions. Compared with available methods through comprehensive simulation studies, the proposed method shows higher power and lower error. The innovative method is applied to clinical pouchitis data as well and interesting results are obtained.

Keywords: high-dimension, mediation analysis, information, nonparametric, microbiome, host genome

OPEN ACCESS

Edited by:
Mariza De Andrade, Mayo Clinic, United States

Reviewed by:
Jaya M. Satagopan, Rutgers, The State University of New Jersey, United States
Brandon Jason Coombes, Mayo Clinic, United States

***Correspondence:**
Lingling An
anling@email.arizona.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 July 2019*
Accepted: *10 February 2020*
Published: *13 March 2020*

Citation:
Carter KM, Lu M, Jiang H and An L (2020) An Information-Based Approach for Mediation Analysis on High-Dimensional Metagenomic Data. Front. Genet. 11:148. doi: 10.3389/fgene.2020.00148

INTRODUCTION

Humans maintain a close symbiotic relationship with trillions of microorganisms that live upon and within their bodies. The human body relies on assorted communities of microbes to develop bodily functions such as metabolism and immune response as well as to protect the body from infections from harmful pathogens. Researchers have begun to recognize the importance of the interactions between host and microbiota and how they may impact human health. In particular, studying this interaction has become a key topic in numerous fields of research such as immunology (Rogers and Wesselingh, 2016; Rooks and Garret, 2016), oncology (Taur and Parmer, 2016), and metabolomics (Rostami et al., 2015; Galla et al., 2017; Kurilshikov et al., 2017). The current Integrative Human Microbiome Project (IHMP) aims to record behavior over time for host biology and the metagenome for the onset of Inflammatory Bowel Disease and Type 2 Diabetes as well as for neonatal development. With progressively more data available, a growing research interest has

emerged for integrative analysis of multiple omics data, for example, host transcriptome and human microbiome data.

One popular approach for integrating multiple omics datasets is mediation analysis. A mediation model aims to extract the mechanisms by which an exposure impacts the outcome variable by considering a set of potential variables which may mediate the effect. Identifying these mechanisms is a vital step in developing effective medication and therapy as well. In particular, the microbial community could be easier to manipulate using antibiotics and probiotics.

Simple mediation models with only one exposure and one mediator have been widely used in psychology for several decades (MacKinnon et al., 2006; Agler and De Boeck, 2017), with most recent notable development focused on models with multiple mediator variables (Daniel et al., 2015). However, the application of mediation models for biological data has introduced additional challenges, including the difficulty of incorporating multiple, high dimensional omics datasets with varying data structures. In this research, we aim to develop a nonparametric framework for mediation analysis to avoid the assumptions and pitfalls of current mediation models.

MATERIALS AND METHODS

Background

A simple mediation model aims to explain the mechanisms that underlay the relationship between an exposure variable (X) and a response variable (Y), by considering a tertiary mediator variable (M) which may mediate the effect of the exposure on the response (**Figure 1**). The total effect of the exposure variable can be decomposed into the *direct effect*, effect from exposure to response directly, and the *indirect effect*, effect of the exposure which is mediated by the mediator variable.

A mediation model is most commonly examined parametrically utilizing a linear structural equation model (LSEM):

$$Y = \gamma' X + \varepsilon \tag{1}$$

$$M = \alpha X + \varepsilon_M \tag{2}$$

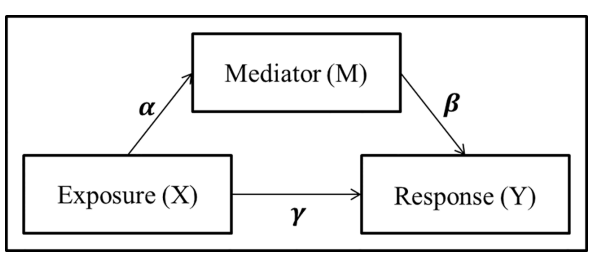


FIGURE 1 | Panel model representation of a mediation model with a single exposure *X*, single mediator *M*, and single response *Y*, and the associated coefficients used for linear structural equation modeling.

$$Y = \gamma X + \beta M + \varepsilon_Y$$

$$= (\gamma + \alpha\beta)X + \beta\varepsilon_M + \varepsilon_Y \tag{3}$$

where γ' and γ represent the total effect and direct effect, respectively. Baron and Kenny (1986) proposed to detect whether an indirect effect exist by testing either the product $\alpha\beta = 0$ or the difference between the total and direct effects $\gamma' - \gamma = 0$. In addition to the traditional mediation assumptions of causal direction (i.e., additive effects and no unmeasured confounders or sequential confounders) (MacKinnon et al., 2006; Vanderwheele and Vansteelandt, 2014; Preacher, 2015), the LSEM approach requires standard regression assumptions such as linearity, no collinearity, known link function, exponential distribution of the error term, and sample size larger than parameter space. While the LSEM structure has seen widespread use and success in psychology applications where mediation analysis includes a single mediator and continuous exposure variables, many of these assumptions are violated in the context of genomics and metagenomics studies with counts data.

In response to these challenges, new statistical methods have been developed in the last few decades in an attempt to apply mediation modeling approaches for neural and biological data. Boca et al. (2014) constructed a distribution of the correlation between parameters by permuting the outcome in each of the LSEM equations. Huang and Pan (2015) developed a Monte-Carlo procedure to evaluate the mediation effect of high-dimentional continuous mediators. Huang et al. (2015) performed an omnibus test by comparing L_1 normalized terms from three logistic regression models based on the structural equations model. Kim et al. (2016) and Nguyen et al. (2016) utilized binary exposure to generate natural direct and indirect effect measures *via* expectation differences. Zhang et al. (2016) used minimax concave penalty regularized logistic regression models to estimate β effect (in eq(3)). Recently, Sohn and Li (2019) proposed a causal composition mediation model (CCMM) specifically for microbiome mediators which utilized a bootstrap covariance matrix to perform log-contrast compositional regression. While these approaches may avoid concerns associated with the $n << p$ paradigm (i.e., sample size is smaller than the parameter space), they often require a single exposure variable and a linear relationship between parameters. Many additionally enforce certain data type such as binary exposures or continuous responses.

In this research, we aim to evaluate the presence of indirect effects by developing a nonparametric framework based on information transfer. While applications of information theory in a biological context have been seldom, it has achieved some success in feature selection for gene expression data (Meyer et al., 2008; Radovic et al., 2017). Recent advances in this field include alternatives for finding relative contribution of variables using entropy methods. Radovic et al. (2017) approached this problem by introducing a penalty term for mutual information shared between selected variables. Liu et al. (2016) assigned a measure of feature quality by comparing conditional information of a variable on an outcome conditioned upon k-nearest-neighbor variables. By utilizing information-based methods, in our research, there is no need to assume underlying distributions

or data types of genomic/metagenomic data, or response variable (e.g., clinical outcome) while nonlinear or non-additive relationships between variables can be explored.

Methods

Recent research has discovered that the abundance and diversity of the microbiome have an impact on the expression of human genes (Blekhman et al., 2015; Bonder et al., 2016; Davenport, 2017). In this study, we will focus on treating microbes as mediators for host genes. However, the proposed method itself is very general and can be applied in other types of studies, e.g., genomic or epigenomic study, or even studies in other fields.

To discover which microbial taxa mediate the effect of gene expression on a clinical outcome, we propose a nonparametric framework based on information theory feature reduction techniques, termed as Nonparametric Entropy Mediation (NPEM). Information theory compares joint distributions of two or more variables with the marginal distributions of subsets to measure association between variables. This can capture nonlinear and non-additive associations by observing changes in distribution of the outcome as compared to distance based and regression modeling approaches which can only capture linear association with the outcome (Roulston, 1999). The information can be measured using Shannon Entropy and Mutual Information (MI) (Shannon, 1949). Shannon entropy represents the uncertainty, potential information, from a discrete random variable or random vector, and is defined as amount of information produced by a stochastic process:

$$H(X) = -\Sigma_{x\in X} p(x)\log p(x), \tag{4}$$

where $p(x)$ represents the probability of observing $X = x$ (if the variable is continuous, this definition is redefined by using the integral across the domain for continuous density functions instead of the summation across the domain of events). Shannon entropy of a multivariate process between two variables X and Y can be calculated using joint Shannon entropy:

$$H(X,Y) = -\Sigma_{x\in X}\Sigma_{y\in Y} p(x,y)\log p(x,y), \tag{5}$$

where $p(x,y)$ represents the probability of observing $X = x$ and $Y = y$ (note: the notations X and Y here are just two common variables, different from the notations in the LSEM in *Background*).

Mutual information (MI) is defined as the overlap of information produced by multiple stochastic processes:

$$MI(X,Y) = H(Y) + H(X) - H(X,Y)$$

$$= \Sigma_{x\in X}\Sigma_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{6}$$

Mutual information can be used as a measure of dependency between the variables in a multivariate stochastic process. If the included variables are independent, the information metric is zero.

To capture the unique mutual information from a variable X, we additionally define the *contributed information* to be the mutual information of one variable given a set of measured variables ($\boldsymbol{W}$):

$$C(X,Y,\boldsymbol{W}) = MI(X,Y) - \Sigma_{w\in W}\frac{MI(X,w)}{\|\boldsymbol{W}\|^2} \tag{7}$$

To investigate the mediated relationship between host gene expression and a clinical outcome, we propose to construct the mediation model as a multivariate stochastic process generating the set of I genes ($\boldsymbol{X} = \{X_1,\ldots,X_I\}$), the set of J microbial taxa ($\boldsymbol{M} = \{M_1,\ldots,M_J\}$), and a clinical outcome Y (throughout the text of this paper we use bold symbols to represent sets of variables). If we maintain the causal direction and no intermediary confounding assumptions, we can examine the relationship between variables using the mutual information between variables from the stochastic processes. To mimic current LSEM structure, we define γ' as a label of relationship between $\boldsymbol{X}$ and Y, α as the relationship between $\boldsymbol{X}$ and $\boldsymbol{M}$, and β as the relationship between $\boldsymbol{M}$ and Y when $\boldsymbol{X}$ is also included. Thus, we use these labels to represent the relationships between the variables based on the theory information in **Figure 2**.

Consider the β effect from $\boldsymbol{M}$ to Y as the overlap in information contained by $\boldsymbol{M}$ and Y, then it can be decomposed into β_1 representing the overlap of α and β, and β_2 representing the unique information from $\boldsymbol{M}$ as shown in **Figure 2** such that $\beta = \beta_1 + \beta_2$. Note that β_2 represents the value $\beta\epsilon_M$ in equation (3). If $\beta_2 \neq 0$, then it follows $\beta \neq 0$. Consider two possible outcomes when $\beta_2 = 0$: 1) if $\beta_1 = 0$ and $\beta_2 = 0$, then $\boldsymbol{M}$ does not offer any information about Y and there is no mediation effect. This is equivalent to $\beta = 0$ and by extension $\alpha\beta = 0$ in the LSEM framework; 2) if $\beta_1 \neq 0$ and $\beta_2 = 0$, all information $\boldsymbol{M}$ provides

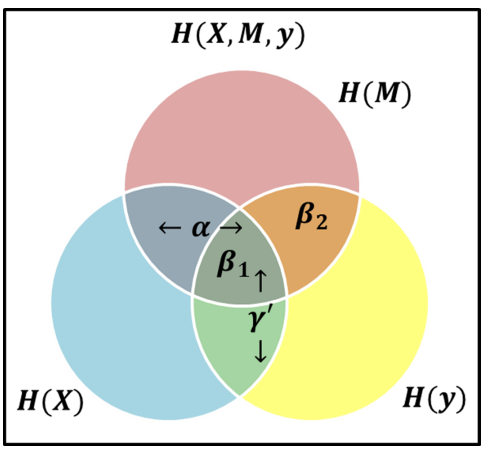


FIGURE 2 | Venn Diagram visual representation of information content and the areas representative of model effects, α represents the relationship (i.e., intersection) between the exposure (blue circle) and mediator (red circle), β_1 is the relationship between all three variables, i.e., at the intersection of three circles, while β_2 represents the relationship between the mediator and response (yellow circle) excluding the exposure, i.e., the area of intersection of red and yellow circles, but not in blue, and γ' is the overlap of blue and yellow circles, representing the total effect.

about Y is also contained in $\boldsymbol{X}$. Due to perfect collinearity, no conclusion can be drawn about the existence of mediation effects. For the purposes of our study, we will consider this scenario as not a mediation effect. Thus, the overlap of all variables is not sufficient and any scenario where $\beta_2 = 0$ would not be considered a mediation effect. The existence of mediation effects can be captured by measuring α and β_2. The two relationships α and β_2 as shown in **Figure 2** can be expressed in terms of mutual information as $MI(\boldsymbol{X}, \boldsymbol{M})$, and $MI(\boldsymbol{M}, Y)$ respectively.

In order to capture the effect of each gene or each taxon individually, we additionally consider collinearity between the variables. We will use contributed information to measure the relationship between gene i and taxon j, $\alpha_{i,j}$, as $C(X_i, M_j, \boldsymbol{S})$, and the relationship between taxon j and the response (for the purpose of explanation we use one clinical response variable) Y, β_2, as $C(M_j, Y, \boldsymbol{T})$, where $\boldsymbol{S}$ and $\boldsymbol{T}$ represent a subset of other genes and other microbial taxa, respectively.

To non-parametrically estimate the mutual information and contributed information metrics, we employ kernel density estimation to approximate the distribution of each variable or a set of variables. To allow for varying data types in a joint distribution, we employ kernel product estimation developed by Li and Racine (2003). The choice of kernel will depend on the structure of the data. For continuous data, the distribution will be approximated using a second order Gaussian kernel, which is a common choice due to its smoothness and an ideal choice when integration is required. Distributions of discrete data will be approximated using an Aitchison-Aitken kernel to handle discrete entry frequencies. To avoid overfitting, bandwidths for kernels are approximated using Silverman's Rule of Thumb (Silverman, 1986). To get an accurate density estimator we only need to know the data type but not the shape.

In high resolution sequencing studies, limited genetic material and PCR amplification biases can lead to many OTUs (operational taxonomic units) with zero count, even when those taxa exist within a subject's gut microbiota. However, a concentration of counts at zero can lead to a problem when estimating the distribution using a Gaussian kernel density estimator. Most notably, the decreased variance can lead to smaller estimates for the kernel bandwidth. We propose two approaches for mediation testing using mutual information. In the simplest case, we use a single Guassian kernel to estimate the distribution of OTU abundance and to calculate the contributed information. We refer to this single kernel approach as a univariate entropy measure. To better represent the microbiome data and to avoid some of the potential pitfalls of kernel density estimation, we propose a bivariate approach which decomposes the microbiome data into two parts: presence-absence represented by an Aithison-Aitken kernel and nonzero counts represented by a Gaussian kernel. Contributed information metrics can be calculated separately for both presence–absence and nonzero counts, providing two measurements for each mediator. We refer to this two-kernel approach as a bivariate entropy measure.

Univariate Entropy Measure

When calculating mutual information, theoretically, the information metric should be zero if the variables are independent; however finite sample sizes and bandwidth approximation for the kernel density estimates may lead to a bias in the observed information. Out of a large number of taxa in a study, generally only some of them play mediating effect. Under this very general assumption, a vast majority of the signals observed are due to this bias effect. Therefore, we can search for information metrics which are substantially higher than the expected bias, as this indicates a true relationship between variables. For a particular taxon (j) to be a mediating taxon, there must be significant relationships from at least one gene through it to the response. Just like the regression model in Eq (2) where all exposure variables $\boldsymbol{X}$ are included for each mediator variable M_j, $\alpha_{i,j}$ (representing the relationship between the exposure variable X_i and mediator M_j) must be evaluated across all exposures simultaneously within each fixed taxon j. For each taxon j the hypotheses are:

$$H_0 : C(X_i, M_j, \boldsymbol{S}) \le \varphi_{\alpha,j}, \forall\, i \in \{1, \ldots, I\} \quad \text{OR} \quad C(M_j, Y, \boldsymbol{T}) \le \varphi_{\beta 2}$$

$$Ha : \exists\, i \in \{1, \ldots, I\} : C(X_i, M_j, \boldsymbol{S}) > \varphi_{\alpha,j} \quad \& \quad C(M_j, Y, \boldsymbol{T}) > \varphi_{\beta 2}$$

The parameters $\varphi_{\alpha,j}$ and $\varphi_{\beta 2}$ represent the expected bias for contributed information with a fixed taxon j and Y respectively. Since the mutual information score should be zero for independent random variables, the bias terms $\varphi_{\alpha,j}$ and $\varphi_{\beta 2}$ are conservatively estimated as the mean contributed information scores for taxon j and currently unselected genes as defined below, respectively:

$$\varphi_{\alpha,j} = \Sigma_{X_i \in (\boldsymbol{X} - \boldsymbol{S})} \frac{C(X_i, M_j, \boldsymbol{S})}{\| (\boldsymbol{X} - \boldsymbol{S}) \|} \tag{8}$$

$$\varphi_{\beta 2} = \Sigma_{M_j \in (\boldsymbol{M} - \boldsymbol{T})} \frac{C(M_j, Y, \boldsymbol{T})}{\| (\boldsymbol{M} - \boldsymbol{T}) \|} \tag{9}$$

where $\boldsymbol{X}$-$\boldsymbol{S}$ represents the set of genes which are currently unselected and $\boldsymbol{M}$-$\boldsymbol{T}$ represents the set of OTUs which are currently unselected. For our definition, both the contributed information and the expected bias depend on the components of set $\boldsymbol{S}$ or $\boldsymbol{T}$. We propose to iteratively select the best predictive genes or taxa based on their contributed information and update $\boldsymbol{S}$ or $\boldsymbol{T}$ respectively after each selection by using a greedy search algorithm. Under this paradigm, we compare the largest contributed information to the average contributed information as defined in equations (8) and (9). This lends itself naturally to outlier detection tests which compare the maximum value to the mean for potential outlier points. Since there could be multiple features which contain true contributed information signals, we opt to use an iterative one-sided Extreme Studentized Deviate (ESD) test (Grubbs, 1950), which was developed for unusually high value detection. We evaluate a series of G statistics (Grubbs, 1950) as follows:

$$G = \frac{C_{(1)}(\ldots) - \overline{C(\ldots)}}{sd(C(\ldots))}$$

where $C_{(1)}$ represents the highest contributed information to be compared, either for the relation between taxon (**j**) and genes, or

for the relation between the outcome and taxa. $\overline{C(\ldots)}$ stands for the average of contributed information and sd represents standard deviation. Under the null hypothesis, the G statistic follows a central t-distribution with degrees of freedom df-2, where df represents the number of remaining unselected features. However, since the contributed information could change at each step, there is still uncertainty on when the hypothesis test should be performed. We propose **Algorithm 1** which performs the hypothesis test at each iteration of the greedy search algorithm (NPEM : UV). To be specific, at each step of the algorithm, the contributed information from each gene to a fixed taxon or from each taxon to the clinical outcome is re-evaluated to identify the most informative feature. The highest value of contributed information is recorded, the hypothesis test is performed, and the selected feature is removed from the set of explanatory variables and added to the set of priors $\boldsymbol{S}$ or $\boldsymbol{T}$. A modified version which performs the hypothesis test after the completion of the greedy search is provided in **Supplementary File** as **Algorithm 1′** (NPEM : UVS). The details and trade-offs of each algorithm are elaborated in the **Supplementary File**.

ALGORITHM 1 | Non-Parametric Entropy Mediation: Univariate Test (NPEM:UV).

Input: $\boldsymbol{A} = \{A_1, A_2, \ldots, A_K\}$: Set of explanatory variables; B: Response variable
1. Initialize an empty set $\boldsymbol{W}$.
2. Evaluate Contributed Information $C_i = C(A_i, B, \boldsymbol{W})$ for each A_i which is not in $\boldsymbol{W}$. When $\boldsymbol{W}$ is empty, $C_i = MI(A_i, B)$.
3. Let $\boldsymbol{C}$ denote the vector of the C_i values, and $C_{(1)}$ denote the largest Contributed Information.
4. Calculate Grubb's ESD Test Statistic: $G = \frac{C_{(1)} - \overline{\boldsymbol{C}}}{sd(\boldsymbol{C})}$, where $\overline{\boldsymbol{C}}$ is the average value and sd represents standard deviation.
5. Perform significance test with the distribution $t_{df\text{-}2}$ to obtain p-value, where df is the length of $\boldsymbol{C}$.
6. If the p-value is below a threshold (e.g., 0.05), move the variable $A_{(1)}$ corresponding to the largest value $C_{(1)}$ into set $\boldsymbol{W}$.
7. Repeat steps 2 through 6 until a specified threshold (e.g. 0.05) is reached or until two or fewer variables remain.
8. For the variables which do not belong to $\boldsymbol{W}$, assign the p-value to be 1.
9. For each response variable, apply FDR correction (Benjamini and Hochberg, 1995) to the p-values of all explanatory variables.

This algorithm is general and can be applied to evaluate the significance of all α and β_2 relationships defined in Methods. For the α relationship, $\boldsymbol{A}$ is the full gene set $\boldsymbol{X}$ and B is an individual microbial taxon (M_j), and the resulting p-value $p_{\alpha,j}$ is the FDR corrected p-values. For the β_2 relationships, $\boldsymbol{A}$ is the set of all microbial taxa $\boldsymbol{M}$ and B is the clinical response (Y). The resulting p-value $p_{\beta,j}$ is FDR corrected. To complete the hypothesis test for mediation effects, we composite the results with conservative measure $p_j = \max(p_{\alpha,j}, p_{\beta,j})$, which represents the final p-value for testing the mediation effect of taxon j.

Bivariate Entropy Measure

When we represent the abundance of each microbial taxon by decomposing the feature into presence-absence and nonzero counts, the contributed information can be calculated for both presence-absence and nonzero counts individually. Our final decision will leverage both contributed information scores. To test whether a relationship is significant or not, we propose a general hypothesis as follows:

$$H_0 : \left\| \vec{C} \right\| \leq \varphi \; vs. \; H_a : \left\| \vec{C} \right\| > \varphi$$

where $\| \vec{C} \|$ represents any norm or distance metric for the vector of two contributed information metrics $\vec{C}$ from zero and nonzero counts. To account for the difference in scale and correlation between presence-absence and nonzero counts, we will utilize Mahalanobis distance (Mahalanbois, 1936):

$$MD\left(\vec{C}\right) = \sqrt{\left(\vec{C} - \vec{\mu}\right)' \Sigma^{-1} \left(\vec{C} - \vec{\mu}\right)},$$

where $\vec{\mu}$ represents the vector of means for $\vec{C}$ and Σ represents the covariance of the two contributed information scores in $\vec{C}$. The Mahalanobis distance is distance metric which projects data along its principal components. Each axis is re-scaled to ensure a mean value of zero and variance of 1. By projecting the two contributed information scores onto their principal components, we no longer need to consider correlation between scores. We can now rewrite our hypothesis using the distance from expected bias:

$$H_0 : MD\left(\vec{C}\right) \leq \varphi \text{ vs. } H_a : MD\left(\vec{C}\right) > \varphi$$

As in the univariate case (i.e., do not separate the zero and nonzeros counts for each taxon) in *Univariate Entropy Measure*, for a particular taxon to be a mediating taxon, there must be a significant mediation structure or bridge from at least one gene and then through the taxon to the clinical response. For each fixed taxon j, the hypotheses are as follows:

$$H_0 : MD\left(\overrightarrow{C_{\alpha,i,j}}\right) \leq \varphi_{\alpha,j}, \forall \; i \in \{1, \ldots, I\} \;\; OR \;\; MD\left(\overrightarrow{C_{\beta_2,j}}\right) \leq \varphi_{\beta_2}$$

$$H_a : \exists i \in \{1, \ldots, I\} : MD\left(\overrightarrow{C_{\alpha,i,j}}\right) > \varphi_{\alpha,j} \;\; \& \;\; MD\left(\overrightarrow{C_{\beta_2,j}}\right) > \varphi_{\beta_2}$$

Since the Mahalanobis projection has two dimensions (i.e., for zero and nonzero parts), we compare the Mahalanobis distance to the Chi-Square distribution with 2 degrees of freedom to identify unusually high contributed information values (De Maesschalck et al., 2000). We provide **Algorithm 2** below which performs the hypothesis test at each iteration of the greedy search algorithm (termed as NPEM : BV). A modified version which performs the hypothesis test after the greedy search algorithm has completed is provided in **Supplementary File** as **Algorithm 2′** (NPEM : BVS). The algorithm follows the same logic as the univariate case, except that we evaluate the contributed information twice, once for the presence-absence data and once for nonzero counts data, with the most informative feature being decided by the largest Mahalanobis distance. The details for obtaining the final p-values are the same as for the univariate test approach.

ALGORITHM 2 | Non-Parametric Entropy Mediation: Bivariate Test (NPEM:BV).

Input: $\boldsymbol{A} = \{A_1, A_2, \ldots, A_K\}$: Set of explanatory variables; B: Response variable
1. Initialize an empty set $\boldsymbol{W}$.
2. For each mediator, decompose into presence-absence and nonzero count (Z, M')
3. Evaluate Contributed Information for both parts (e.g. $\vec{C_i} = \{C_Z = C(A_i, Z, \boldsymbol{W}), C_{M'} = C(A_i, M', \boldsymbol{W})\}$) for each A_i which is not in $\boldsymbol{W}$.
4. Evaluate the Mahalanobis distance for each vector of contributed information scores $\vec{C_i}$.
5. Move variable A_k into set $\boldsymbol{W}$.
6. Calculate the Chi-Square Test Statistic: $\chi^2 = MD(\vec{C}_{(1)})$
7. If the p-value is below a threshold (e.g., 0.05), move the variable $A_{(1)}$ corresponding to the largest Mahalanobis distance $MD(\vec{C}_{(1)})$ into set $\boldsymbol{W}$.
8. Repeat steps 3 through 7 until a specified threshold is reached (e.g. 0.05) or until two or fewer variables remain.
9. For the variables which do not belong to W, assign the p-value to be 1.
10. For each response variable, apply FDR correction to the p-values of all explanatory variables.

Data

Simulation Studies

To evaluate the performance of NPEM, we compare our method to existing methods, a nonparametric permutation test, MedTest (Boca et al., 2014), and a method developed to handle SNP counts data, Integrative Genome Wide Association Study, iGWAS (Huang et al., 2015). We simulate biological data for a dichotomous clinical outcome (e.g., healthy or diseased) under various model settings. Gene expression data was simulated for 300 genes using a normal distribution. The first 150 were generated using a standard deviation of 0.5, and the second half with 2.0. Taxon counts were generated using a negative binomial distribution with excess zeros added, with the probability of excess zeros weighted by the log ratio of abundance to population mean (see the **Supplementary File**). The relationships between variables are presented in **Table 1** below.

Three separate simulation studies are performed to examine the behaviour of NPEM under different scenario settings:

i. The first study investigates the performance of different models with various sample size (40 and 80 per group) and excess zero probabilities at a high level (80%) or low level (50%) for a total of four data scenarios. The signal strength is fixed at 50%, which is defined as follows:

$$\text{signal strength} = \frac{\delta}{\sigma}$$

where δ represents the average difference between healthy and diseased groups and σ represents the standard deviation of the noise.

ii. In practical studies, the signal strength is unlikely to be large for each taxon. We investigate how these methods perform as signal strength decreases by varying signal strength between 50% and 10%. In this simulation, we also vary the excess zero proportions between high (80%) and low (50%), with a fixed sample size 40 per group.

iii. For further investigation, we observe the effects by increasing the over-dispersion of taxon counts. The over-dispersed counts are modeled using a negative binomial model with the dispersion parameter as follows:

$$\kappa = \frac{c}{\sqrt{\lambda + 1}}.$$

where λ represents the mean count and the constant c is set to 1000 for high dispersion and 100 for low dispersion. We fix the sample size to be 40 per group and the excess zero proportion at high (80%) to capture the worse-case. Signal strength ranges from 10% to 50% as in simulation (ii).

For each scenario a total of 20 data sets are generated and evaluated. The results of the simulation studies are presented in *Results*.

TABLE 1 | Existence of relationships for combinations of gene and taxon indices. True mediation effects require γ' (total effect), α, and β_2 relationships. Here taxa 1–10 are the true mediators for genes 1–20, and taxa 151–160 are the mediators for genes 151–170. The rest taxa are not mediators.

	Low expression σ = 0.5		High expression σ = 2	
	Genes 1–20	Genes 21–150	Genes 151–170	Genes 171–300
Taxa 1–10	γ', α, β_2	β_2	γ', β_2	β_2
Taxa 11–20	γ', α		γ'	
Taxa 21–30	γ', β_2	β_2	γ', β_2	β_2
Taxa 31–150	γ'		γ'	
Taxa 151–160	γ', β_2	β_2	γ', α, β_2	β_2
Taxa 161–170	γ'		γ', α	
Taxa 171–180	γ', β_2	β_2	γ', β_2	β_2
Taxa 181–300	γ'		γ'	

Pouchitis Data

Pouchitis, inflammation of a post-operation ileal pouch, affects almost half of all ileal pouch-anal anastomosis recipients, with up to 20% of these patients developing chronic pouchitis. We apply NPEM to pouchitis patient data from Morgan et al. (2015), including host gene expression, microbial abundance, and clinical diagnosis, to investigate the relationship of the host gene expression and microbiome. While extensive research has shown host gene expression and the microbiome can influence pouchitis, the causal mechanisms and interactions are not studied well and the authors only found weak association between host gene expression and the microbiome's effects on the clinical diagnosis.

The clinical data includes samples from 219 patients with information about body location, inflammatory score, antibiotic use, and clinical diagnosis of "No Pouchitis", "Acute Pouchitis", "Chronic Pouchitis", "Crohn's Disease-Like", and "Familial Adenomatous Polyposis". For comparison purposes, we have limited our study to patients with either "No Pouchitis" or "Acute Pouchitis" diagnoses, and no prescribed antibiotics given. This results in an effective sample size of 101 patients. Gene expression data contains 33,297 genes. Transcripts were filtered to remove genes with no annotation, and a log-2 fold change with a conservative cut-off of 0.15 was used to trim the

gene set. After filtering, 1103 genes remained. High throughput next-generation sequencing microbiome abundance data recorded 293 operational taxonomic units (OTUs) at the genus level. OTUs that were absent in over 90% of patients were removed, resulting in 103 OTUs.

RESULTS

Simulation Study Results

With a false positve rate of 5%, NPEM algorithms have higher power than MedTest, while iGWAS fails to discover any significant mediators (**Figure 3**). From the study (i) where the signal strength is high, we find that the UV version of the univariate approach consistently performs the best and the UVS does not perform as well as other NPEM algorithms. Particularly, for a high proportion of zeros and small sample size the UV surpasses the others. As the signal size decreases from 50% to 10% (**Figure 4**), the performance of this univariate test decreases, regardless of the levels of proportion of zeros. However, the bivariate approach maintains better performance. In particular, the single test (BVS) of the bivariate approach is the most consistent and has the highest power when the proportion of zeros in the dataset is high; for a lower proportion of zeros the BV approach is recommended.

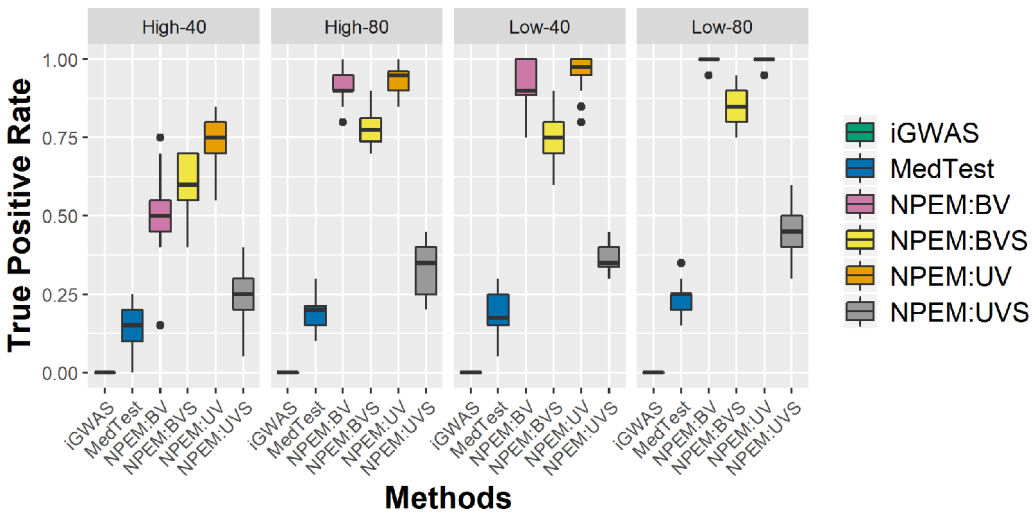


FIGURE 3 | Power plots for simulation studies (i). Sample sizes (40 and 80 per group) and proportions of zero (Low vs. High), with a fixed high signal strength.

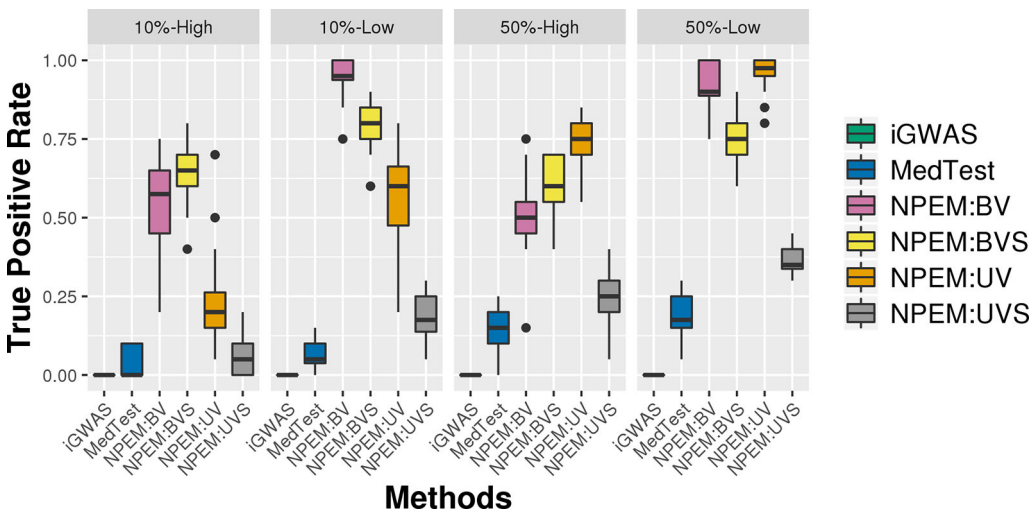


FIGURE 4 | Power plots for simulation studies (ii). Signal strength (50% and 10%) and proportions of zeros (Low vs. High), with a fixed sample size.

For the overdispersion study (i.e., setting iii), the lower the overdispersion, the higher the power (**Figure 5**). The UV approach always outperforms the alternatives when the signal strength is higher, regardless the overdispersion levels; the BVS is always the superior method when the signal size is lower.

For all simulation settings and all methods, the empirical false positive rates are well controlled at pre-specified level. For instance, under simulation setting (i) and using an adjusted p-value cut-off at 0.05, the false positive rates are well controlled (**Figure 6**). The results for settings ii) and iii) are available in the **Supplementary File**.

Pouchitis Study Results

Due to zero proportions ranging from 20% to 90%, moderate sample size, and small expected signals in the pouchitis OTU data, we applied the proposed approach BVS on this dataset. Six mediating OTUs were detected at 5% FDR level and the corresponding genera are summarized in **Table 2**. To visualize the relationship between the detected genera and their significantly

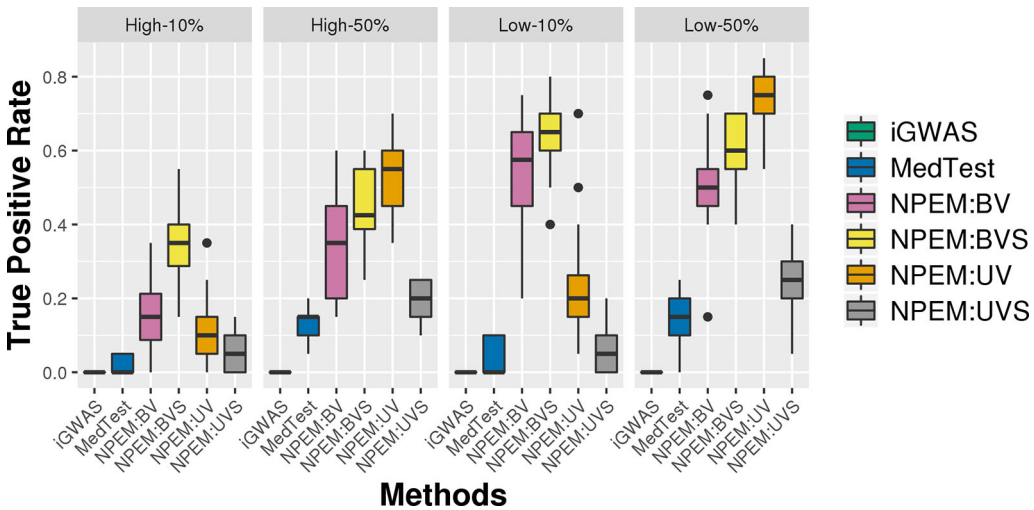


FIGURE 5 | Power plots for simulation studies (iii). Over-dispersion (Low and High) and signal strength (50 and 10%), with a fixed sample size and a fixed proportion of zeros.

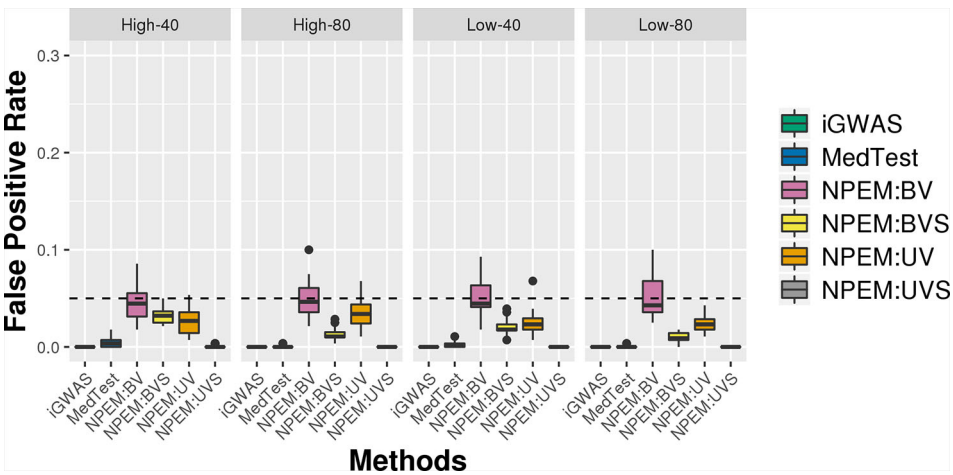


FIGURE 6 | False positive rate plots for simulation studies (i). Sample sizes (40 and 80 per group) and proportions of zeros (Low vs. High), at a fixed signal strength.

TABLE 2 | Top 6 selected Genera with adjusted P-values from NPEM : BVS algorithm.

Genus	Adjusted p-value
Spirochaeta	4.13E-05
Adlercreutzia	1.96E-04
Propionibacterium	2.15E-04
Scardovia	2.86E-03
Stenotrophomonas	8.34E-03
Fusobacterium	8.91E-03

associated genes, a network plot using significant relationships identified by NPEM : BVS is provided in **Figure 7**.

While research on how bacteria impacts the body is still ongoing, the selected microbial genera are well known to be related to intestinal health. Fusobacterium and stenotrophomonas are well known to be pro-inflammatory (Sasaki and Klapproth, 2012; Shaw et al., 2016), while propionibacterium has recently been found to regulate inflammatory response (Ple et al., 2015; Colliou et al., 2017). Fusobacterium and aldercreutzia are also found to relate to the health of the host mucosal wall (Shaw et al., 2016). Degraded mucosal walls may lead to greater risk of infections due to bacteria growing in the folds of the intestinal wall. Scardovia and spirochaeta have been commonly discovered to be associated with ulcerative and ischemic colitis (Lee et al., 1971; Sasaki and Klapproth, 2012; Xun et al., 2018), two of the primary diseases resulting in ileal pouch-anal anastomosis. Though the exact mechanisms are yet understood, these choices correlate to existing findings and suggest further research is necessary.

When looking at the selected genes, we see a few unique patterns. A number of genes, particularly those related to Scardovia and Stenotropohomonas, are only located on the Y chromosome. Patient gender was not included in the provided metadata, so we were not able to test whether this effect is somehow related to gender or the specific gene. Many selected genes are in the Caspases (CASP) or Small Nucleotide RNA C/D Box (SNORD) groups. CASP genes regulate inflammation response (Scott and Saleh, 2007), which is what we expect. The SNORD gene group regulates expression of other gene groups. In particular, recent research has found correlation between SNORD-116 segments and gut metabolism (Qi et al., 2016). These genes may be a prime candidate for future research.

DISCUSSION

In this paper, we propose nonparametric entropy models to discover significant mediation structures for microbial

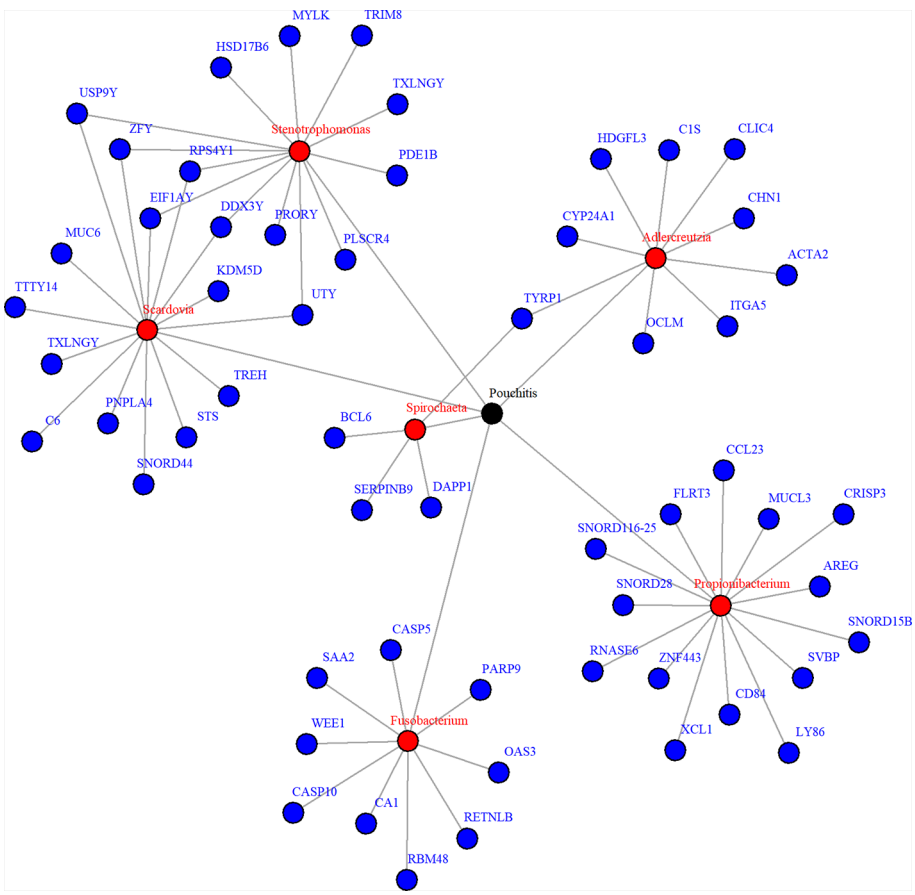


FIGURE 7 | Network plot for significant mediation relationships for detected microbes and associated genes using NPEM : BVS in the Acute Pouchitis study.

mediators. This method is flexible and capable in handling continuous, discrete, and mixed data types for any variable in the model. Though we only discuss continuous and categorical data here, ordinal data may be used in the model by applying a modified Wang-van Ryzen kernel as proposed by Li and Racine (2003) or any other appropriate kernel type. Through simulation studies, we have shown that NPEM outperforms the existing nonparametric test and count-based regression model. In application, our method identifies unique mediation structures undiscovered in the original report relating inflammatory bacteria to host gut health.

The performance of NPEM depends on the data characteristics and selected test statistic. The signal strength in the data is the largest factor separating the performance of the univariate and bivariate options. The bivariate single test (BVS) method is recommended for weak signal size. For the test statistic selection, the poor performance of a singular Grubb's test is expected; the Grubb's test is designed to select singular outliers, thus requires sequential selection. Comparison between the bivariate Chi-Square tests is not straightforward since the correlation structure is re-evaluated at each step of the sequential selection algorithm. The proportion of zeros in the data also affects the test selection. When the excess zero proportion is high, a singular test performs stronger than a sequential test. It is important to recognize that the Mahalanobis distance metric does not consider directionality, and unusually low signals may also be selected. A detailed check may be helpful when the noise signals are large.

The alternative causal compositional mediation model, CCMM (Sohn and Li, 2019) was attempted, however, due to the high proportion of zeros and large number of taxonomic units in our experiment, the CCMM algorithm failed to converge. In toy data experiments with no zero counts, CCMM displays higher power in detecting mediating taxa, however it produces much higher false positive rates for associations between host gene expression and taxonomic abundance since the method does not correct for correlation between exposures. The NPEM methods perform much stronger at detecting the correct associations for this particular path α. CCMM is proposed for continuous response, though theoretically a logit link function could handle a binary response.

The performance of our model may be improved through further tuning. The Gaussian kernel is chosen for approximating log-expression density functions for its smoothness and continuous properties. Other kernel types may provide a more accurate fit of the true distribution. Further research is necessary to conclusively decide on the optimal kernel structures for a given dataset. Additionally, the information metrics may be more accurately estimated by implementing leave-one-out cross-validation at the cost of decreased computation speed. However, this research will be the first research to explore the mediation effect from a brand new point of view, an information-based theory.

DATA AVAILABILITY STATEMENT

16S sequence data for this project was downloaded from Bioproject PRJNA269954 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA269954]. Microarray data are available from GEO as GSE65270 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65270]. Metadata are available at [http://huttenhower.sph.harvard.edu/pouchitis2015].

AUTHOR CONTRIBUTIONS

LA and KC conceived the study. KC designed the methods and algorithms. LA and ML assisted in tuning and critiquing proposed methods. LA and HJ proposed the real data analysis and KC performed the real data analysis. KC drafted the manuscript and all authors edited it.

FUNDING

This work was partially supported by the National Science Foundation [DMS-1222592 to LA]; and the United States Department of Agriculture [ARZT-1360830-H22-138 and ARZT-1361620-H22-149] to L.A.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00148/full#supplementary-material

REFERENCES

Agler, R., and De Boeck, P. (2017). On the Interpretation and Use of Mediation: Multiple Perspectives on Mediation Analysis. *Front. In Psychol.* 8, 1984. doi: 10.3389/fpsyg.2017.01984

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–1182. doi: 10.1037/0022-3514.51.6.1173

Benjamini, Y., and Hochberg, Y. (1995). The False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. (Meth.)* 57, 289–300. doi: 10.2307/2346101

Blekhman, R., Goodrich, J. K., Huan, K., Sun, Q., Bukowski, R., Bell, J. T., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16, 1. doi: 10.1186/s13059-015-0759-1

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C., and Sampson, J. N. (2014). Testing Multiple Biological Mediators Simultaneously. *Bionformatics* 30, 214–220. doi: 10.1093/bioinformatics/btt633

Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Mujagic, Z., Imhann, F., Vila, A. V., et al. (2016). The effect of host genetics on the gut microbiome. *Nat. Genet.* 48, 1407–1412. doi: 10.1038/ng.3663

Colliou, N., Ge, Y., Sahay, B., Gong, M., Zadeh, M., Owen, J., et al. (2017). Commensal Propionibacterium strain UF1 mitigates intestinal inflammation *via* Th17 cell regulation. *J. Clin. Invest.* 127, 3970–3986. doi: 10.1172/JCI95376

Daniel, R. M., De Stalova, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal Mediation Analysis with Multiple Mediators. *Biometrics* 71, 1. doi: 10.1111/biom.12248

Davenport, E. R. (2017). Tooth Be Told, Genetics Influences Oral Microbiome. *Cell Host Microbiome* 22, 251–253. doi: 10.1016/j.chom.2017.08.018

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 50, 1–18. doi: 10.1016/s0169-7439(99)00047-7

Galla, S., Chakrabory, S., Mell, B., Vijay-Kumar, M., and Joe, B. (2017). Micorobiotal-Host Interactions and Hypterension. *Physiology* 32, 224–233. doi: 10.1152/physiol.00003.2017

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat* 21, 27–58. doi: 10.1214/aoms/1177729885

Huang, Y., and Pan, W. (2015). Hypothesis Test of Mediation Effect in Causal Mediation Model with High Dimensional Continuous Mediators. *Biometrics* 72, 2. doi: 10.1111/biom.12421

Huang, Y., Liang, L., Moffatt, M. F., Cookson, W. O. C. M., and Lin, X. (2015). iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genet. Epidemiol.* 39, 5. doi: 10.1002/gepi.21905

Kim, C., Daniels, M. J., Marcus, B. H., and Roy, J. A. (2016). A Framework for Bayesian Nonparametric Inference for Causal Effects of Mediation. *Biometrics* 73, 2. doi: 10.1111/biom.12575

Kurilshikov, A., Wijmenga, C., Fu, J., and Zhernakova, A. (2017). Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends In Immunol.* 38, 633–647. doi: 10.1016/j.it.2017.06.003

Lee, F. D., Kraszewski, A., Gordon, J., Howie, J. G. R., McSeveney, D., and Harland, W. A. (1971). Intenstinal spirochaetosis. *Gut* 12, 126–133. doi: 10.1136/gut.12.2.126

Li, Q., and Racine, J. (2003). Nonparametric estimation of distribution with categorical and continuous data. *J. Multivariate Anal.* 86, 266–292. doi: 10.1016/S0047-259X(02)00025-8

Liu, J., Lin, Y., Lin, M., Shunxiang, W., and Zhang, J. (2016). Feature selection based on quality of information. *Neurocomputing* 255, 11–22. doi: 10.1016/j.neucom.2016.11.001

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2006). Mediation Analysis. *Annu. Rev. Psychol.* 58, 593–614. doi: 10.1146/annurev.pscyh.58.110405.085542

Mahalanbois, P. C. (1936). On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, 249–55.

Meyer, P. E., Schretter, C., and Bontempi, G. (2008). Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity. *IEEE J. Sel. Topics In Signal Process.* 2, 261–274. doi: 10.1109/JSTSP.2008.923858

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Gegome Biol.* 16, 67. doi: 10.1186/s13059-015-0637-x

Nguyen, T. Q., Webb-Vargas, Y., Koning, I. M., and Stuart, E. A. (2016). Causal Mediation Analysis with a Binary Outcome and Multiple Continuous or Ordinal Mediators: Simulations and Application to an Alcohol Intervention. *Struct. Equ. Model.* 23, 3. doi: 10.1080/10705511.2015.1062730

Ple, C., Richoux, R., Jardin, J., Nurdin, M., Briard-Bion, V., Parayre, S., et al. (2015). Single-strain starter experimental cheese reveals anti-inflammatory effect of Propionibacterium freudenreichii CIRM BIA 129 in TNBS-colitis model. *J. Funct. Foods* 18, 575–585. doi: 10.1016/j.jff.2015.08.015

Preacher, K. J. (2015). Advances in Mediation Analysis: A Survey and Synthesis of New Developments. *Annu. Rev. Psychol.* 66, 825–852. doi: 10.1146/annurev-psych-010814-015258

Qi, Y., Purtell, L., Fu, M., Lee, N. J., Aepler, J., Zhang, L., et al. (2016). SNORD116 is critical in the regulation of food intake and body weight. *Sci. Rep.* 6, 1. doi: 10.1038/srep18614

Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinf.* 18, 1. doi: 10.1186/s12859-016-1423-9

Rogers, G. B., and Wesselingh, S. (2016). Precision respiratory medicine and the microbiome. *Lancet Respir. Med.* 4, 73–82. doi: 10.1016/S2213-2600(15)00476-2

Rooks, M. G., and Garret, W. S. (2016). Gut microbiota, metabolites, and host immunity. *Immunology* 16, 341–352. doi: 10.1038/nri.2016.42

Rostami, N. M., Ishaq, S., Al Dulaimi, D., Zali, M. R., and Rostami, K. (2015). The Role of Infectious Mediators and Gut Microbiome in the Pathogenesis of Celiac Disease. *Arch. Iran. Med.* 18, 244–249. doi: 015184/AIM.0010

Roulston, M. S. (1999). Estimating Errors on Measured Entropy and Mutual Information. *Phys. D: Nonlinear Phenom.* 125, 285–294. doi: 10.1016/S0167-2789(98)00269-3

Sasaki, M., and Klapproth, J. A. (2012). The Role of Bacteria in the Pathogenesis of Ulcerative Colitis. *J. Signal Transduct.* 2012, 704953 doi: 10.1155/2012/704953

Scott, A. M., and Saleh, M. (2007). The inflammatory caspases: guardians against infections and sepsis. *Cell Death Diff.* 14, 23–31. doi: 10.1038/sj.cdd.4402026

Shannon, C. E. (1949). Communication in the Presence of Noise. *Proc. IRE* 37, 1. doi: 10.1109/JRPROC.1949.232969

Shaw, K. A., Bertha, M., Hofmekler, T., Chopra, P., Vatanen, T., Srivatsa, A., et al. (2016). Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med.* 8, 1. doi: 10.1186/s13073-016-0331-y

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Density Estimation Stat Data Anal.* doi: 10.1201/9781315140919

Sohn, M., and Li, H. (2019). Compositional Mediation Analysis for Microbiome Studies. *Ann. Appl. Stat* 13, 661–681. doi: 10.1214/18-AOAS1210

Taur, Y., and Parmer, E. (2016). Microbiome mediation of infections in the cancer setting. *Genome Med.* 8, 40. doi: 10.1186/s13073-016-0306-z

Vanderwheele, T. J., and Vansteelandt, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiol. Method* 2, 1. doi: 10.1515/em-2012-0010

Xun, Z., Zhang, Q., Xu, T., Chen, N., and Chen, F. (2018). Dysbiosis and Ecotypes of the Salivary Microbiome Associated With Inflammatory Bowel Diseases and the Assistance in Diagnosis of Diseases Using Oral Bacterial Profiles. *Front. In Microbiol.* 9, 1136. doi: 10.3389/fmicb.2018.01136

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., et al. (2016). Estimating and Testing High-dimensional Mediation Effects in Epigenetic Studies. *Bioinformatics* 32, 3150–3154. doi: 10.1093/bioinformatics/btw351

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.

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: info@frontiersin.org | +41 21 510 17 00

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