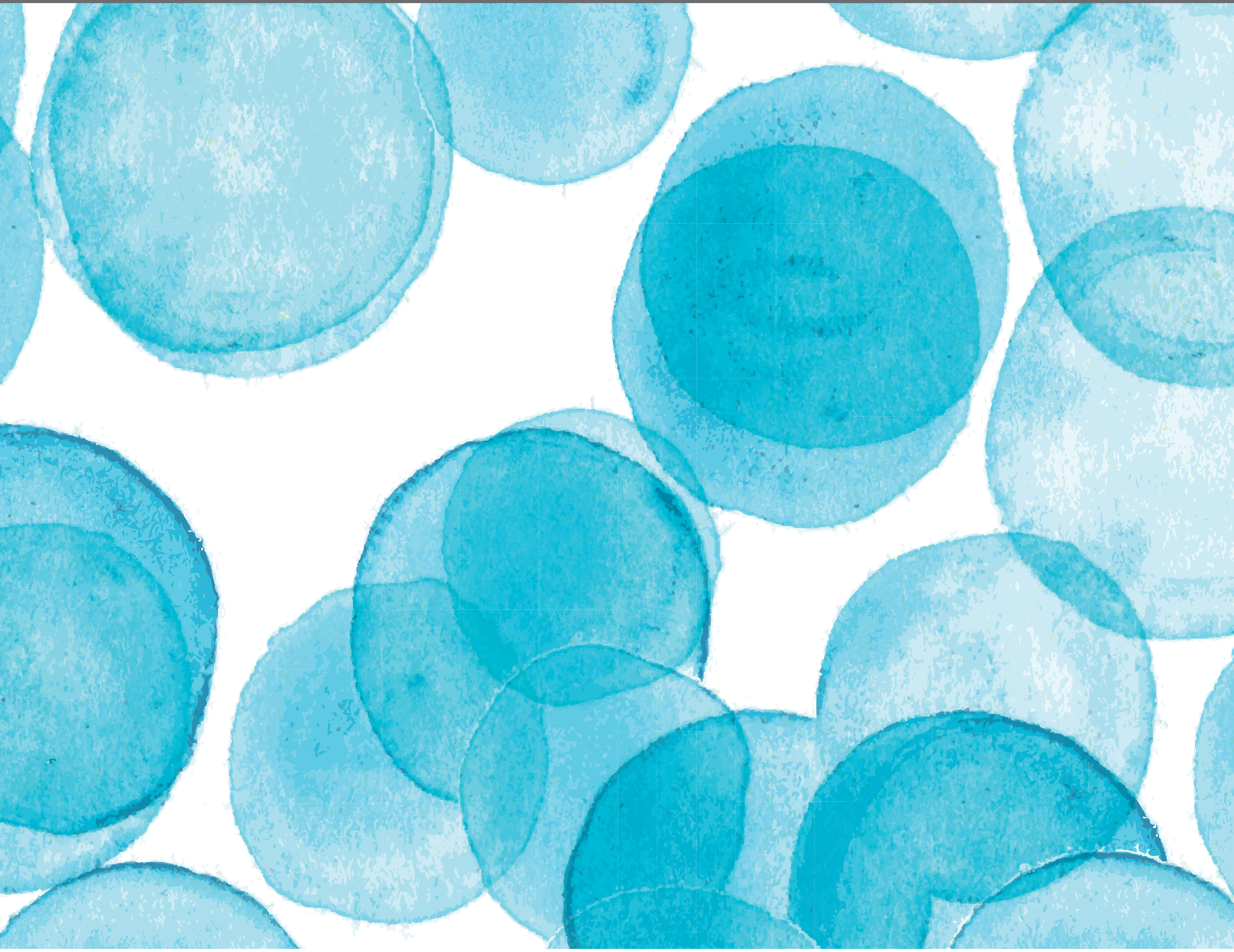
NOVEL APPROACHES IN MICROBIOME ANALYSES AND DATA VISUALIZATION

EDITED BY: Jessica Galloway-Peña and Michele Guindani
PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-653-6
DOI 10.3389/978-2-88945-653-6

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

NOVEL APPROACHES IN MICROBIOME ANALYSES AND DATA VISUALIZATION

Topic Editors:
Jessica Galloway-Peña, MD Anderson Cancer Center, United States
Michele Guindani, University of California, Irvine, United States

High-throughput sequencing technologies are widely used to study microbial ecology across species and habitats in order to understand the impacts of microbial communities on host health, metabolism, and the environment. Due to the dynamic nature of microbial communities, longitudinal microbiome analyses play an essential role in these types of investigations. Key questions in microbiome studies aim at identifying specific microbial taxa, enterotypes, genes, or metabolites associated with specific outcomes, as well as potential factors that influence microbial communities.

However, the characteristics of microbiome data, such as sparsity and skewedness, combined with the nature of data collection, reflected often as uneven sampling or missing data, make commonly employed statistical approaches to handle repeated measures in longitudinal studies inadequate. Therefore, many researchers have begun to investigate methods that could improve incorporating these features when studying clinical, host, metabolic, or environmental associations with longitudinal microbiome data.

In addition to the inferential aspect, it is also becoming apparent that visualization of high dimensional data in a way which is both intelligible and comprehensive is another difficult challenge that microbiome researchers face. Visualization is crucial in both the analysis and understanding of metagenomic data. Researchers must create clear graphic representations that give biological insight without being overly complicated. Thus, this Research Topic seeks to both review and provide novels approaches that are being developed to integrate microbiome data and complex metadata into meaningful mathematical, statistical and computational models. We believe this topic is fundamental to understanding the importance of microbial communities and provides a useful reference for other investigators approaching the field.

Citation: Galloway-Peña, J., Guindani, M., eds. (2019). Novel Approaches in Microbiome Analyses and Data Visualization. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-653-6

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 27 September 2018
doi: 10.3389/fmicb.2018.02274

Editorial: Novel Approaches in Microbiome Analyses and Data Visualization

Jessica Galloway-Peña[1,2]* *and Michele Guindani*[3]

[1] Department of Genomic Medicine, MD Anderson Cancer Center, Houston, TX, United States, [2] Department of Infectious Diseases, Infection Control, and Employee Health, MD Anderson Cancer Center, Houston, TX, United States, [3] Department of Statistics, University of California, Irvine, Irvine, CA, United States

Keywords: microbiome, data visualization, statistics, longitudinal data analysis, metagenomics, bioinformatics

Editorial on the Research Topic

Novel Approaches in Microbiome Analyses and Data Visualization

Next generation sequencing technologies have allowed the study of microbial ecosystems at previously unseen depths. In both ecology and human biology, there is a pressing quest to advance our understanding of how microbial communities impact their host and their environment. In particular, the majority of microbiome studies are aimed at identifying specific microbial taxa, community profiles, genes, or metabolites which may be predictive of specific outcomes, functions, or disease states. However, due to the complexity of microbiome data, the statistical and computational analysis of these data present many challenges which may affect the validity of commonly employed methods. Therefore, despite the fact that microbiome and bioinformatic researchers often use widely accepted pipelines, the field remains wide open for improvement. In this Research Topic, a few researchers have responded to the task of reviewing or describing novel methodologies aimed at tackling the challenges of microbiome data and the respective metadata. Only with the development of improved statistical and computational models can one really hope to exploit microbiome based research to understand biological mechanisms, identify biomarkers of disease, or delineate microbial interactions with their environment.

The largest challenge investigators face in developing statistical approaches to study microbiome data is considering all of the constraints of microbiome data fully. Multiple researchers support regarding microbiome data as compositional, meaning the data are usually described as relative quantitative descriptions as parts of some whole, such as proportions or relative abundance. Of course, this view is also partial since important information may be lost when adopting a compositional perspective. However, intrinsic complications arise among commonly employed techniques if microbiome data are examined using a non-compositional paradigm. Gloor et al. reviews a number of recently proposed compositional data analyses methods for microbiome data, and provides some caveats against a naïve use of statistical models if the data are not treated as compositional.

One common problem among compositional microbiome data is that it is sparse and zero-inflated. This compositional bias leads to false positives as well as underpowered statistical associations when conducting multiple comparisons. A common strategy to handle excess zeros is to add a small number called a pseudo count. Kaul et al. propose a novel method (ANCOM-II) for handling zeros in microbiome data by first identifying the types of zeros in your data, then comparing the abundance of taxa relative to a background or reference value which is present in all specimens. Simulations of the authors' methodology show improved control for false discovery rate and higher statistical power compared to pseudo-counts. Another dilemma is that rare and low abundant taxa naturally exist among microbiomes. Karpinets et al. attempt to reduce the burden of filtering for the rare OTUs and overcome the difficulty of compositionality by treating the OTUs

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Spyridon Ntougias,
Democritus University of Thrace,
Greece

***Correspondence:**
Jessica Galloway-Peña
jrgalloway@mdanderson.org

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

Received: 02 August 2018
Accepted: 06 September 2018
Published: 27 September 2018

Citation:
Galloway-Peña J and Guindani M
(2018) Editorial: Novel Approaches in
Microbiome Analyses and Data
Visualization. Front. Microbiol. 9:2274.
doi: 10.3389/fmicb.2018.02274

as qualitative variables. They explore the biological role of the rare low abundance OTUs and analyze them by using association networks (Anets) and show Anets have the potential to serve as a unsupervised methodology for linking rare OTUs to associations with environment or phenotypes.

Many methodologies naïvely separate themselves from the biologic aspect of the data inasmuch that these are complex interacting ecosystems with intertwined metabolic pathways, different rates of growth etc. Pinto et al. construct an ordinary differential equation (ODE) -based kinetic model incorporating microbial growth equations and metabolic interactions among bacteria using experimental data from gut microbe cultures. Their model accurately predicted bacterial abundance as well as metabolite consumption and production in a bioreactor experiment.

Furthermore, many approaches do not take into consideration phylogeny or relatedness of the organisms in order to make associations. Zhai et al. suggest a variance component selection scheme, or VC-lasso, for sparse and high-dimensional taxonomic data analysis. They disperse individual OTUs into clusters at phylogenetic levels, and translate the phylogenetic distance information to kernel matrices, where they treat the taxonomic clusters as multiple random effects in a variance component model. Similarly, Xiao et al. also develop a methodology for capturing clustered microbiome signals dependent on phylogeny. "glmmTree" is their novel prediction method based on a generalized linear mixed model, which captures clustered microbiome signals. In this framework, the effects of the bacterial taxa are modeled as random with the correlation structure dependent on a phylogenetic tree, whereas the effects of predictive variables are treated as fixed. Another conundrum is the concern that methodologies based on binning mapped sequences can still be riddled with error due to subpar databases. Currently OTU binning is the well accepted methodology, but group specific signatures can be just as important for biomarker discovery or disease association. Wang et al. recommend using K-mers which provides an alignment free method to characterize microbial communities.

There is a definite shortage of visualization or web based tools that support the integration of taxonomic and functional profiles. BURRITO, described by McNally et al. is a web based tool for interactive visualization of microbiome multi-omic data combined with taxonomic and functional information. BURRITO visualizes the taxonomic and functional compositions of multiple samples and underlines relationships between taxa and function. Baksi et al. present a web based framework called "TIME" ("Temporal Insights into Microbial Ecology"). TIME allows for predicting taxa that might have a higher influence on community structure in different conditions.

As substantial variability in microbiota communities can be seen across subjects, and across time, the improvement of longitudinal study design, and causal models is paramount to associate a dynamic ecosystem with complicated environmental and host factors. Many of the papers in this Research Topic offer methods which address different issues that arise when handling longitudinal data. The previously discussed web based application, TIME, was developed specifically to identify potential taxonomic markers from time series data (Baksi et al.). In this program, longitudinal time points, and respective metadata can be used to visualize temporal variations. Lee and Sison-Mangus developed a Bayesian semiparametric generalized linear regression model to investigate the effects of physical and biological variables the abundance of microbes. This model allows for borrowing information across OTUs, across samples and across time points. Shields-Cutler et al. introduce splinectomeR, an R package that uses smoothing splines to summarize categorical variables for hypothesis testing in longitudinal microbiome studies. Lastly, Wagner et al. propose the use of a bi-exponential function to summarize and compare diversity curves over time using hierarchical modeling. This approach accounts for repeated measures on each subject in order to compare and model alpha diversity indices over time.

Together, these original research articles and reviews emphasize the difficulties faced when analyzing microbiome data and the shortcomings of current statistical, computational and visualization tactics. Currently, many researchers perform all of their own coding and individual analyses without well-defined descriptions of their methods or sharing of analysis pipelines between laboratories. Thus, there is a pressing need for consistent and harmonious data analysis procedures. As a scientific community, microbiome researchers should not be content with "status quo" when it comes to widely accepted practices in microbiome analyses as they have many faults and limitations. In the modern era of data sharing and web-based tools scientists should be working together to compare results between sites and cohorts, improve current techniques, as well as validate methodologies. Only when the research community ensures that novel approaches hold up to independent validation across populations can we truly develop a microbiome analysis paradigm which allows for reliable reproducibility of findings across multiple institutions all over the world.

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 assistance in putting together this research topic collection. We thank Samuel S. Shelburne, III for his constructive comments and feedback.

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.

MINI REVIEW
published: 15 November 2017
doi: 10.3389/fmicb.2017.02224

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1]*, *Jean M. Macklaim*[1], *Vera Pawlowsky-Glahn*[2] and *Juan J. Egozcue*[3]

[1] *Department of Biochemistry, University of Western Ontario, London, ON, Canada,* [2] *Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain,* [3] *Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain*

Datasets collected by high-throughput sequencing (HTS) of 16S rRNA gene amplimers, metagenomes or metatranscriptomes are commonplace and being used to study human disease states, ecological differences between sites, and the built environment. There is increasing awareness that microbiome datasets generated by HTS are compositional because they have an arbitrary total imposed by the instrument. However, many investigators are either unaware of this or assume specific properties of the compositional data. The purpose of this review is to alert investigators to the dangers inherent in ignoring the compositional nature of the data, and point out that HTS datasets derived from microbiome studies can and should be treated as compositions at all stages of analysis. We briefly introduce compositional data, illustrate the pathologies that occur when compositional data are analyzed inappropriately, and finally give guidance and point to resources and examples for the analysis of microbiome datasets using compositional data analysis.

Keywords: microbiota, compositional data, high-throughput sequencing, correlation, Bayesian estimation, count normalization, relative abundance

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Ionas Erb,
Centre for Genomic Regulation, Spain
Jennifer Stearns,
McMaster University, Canada

****Correspondence:***
Gregory B. Gloor
ggloor@uwo.com

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

Received: *10 July 2017*
Accepted: *30 October 2017*
Published: *15 November 2017*

Citation:
Gloor GB, Macklaim JM, Pawlowsky-Glahn V and Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8:2224. doi: 10.3389/fmicb.2017.02224

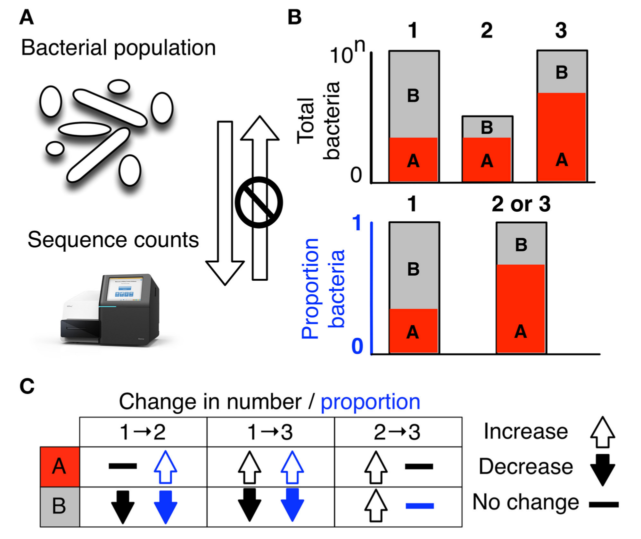
1. INTRODUCTION

The collection and analysis of microbiome datasets presents many challenges in the study design, sample collection, storage, and sequencing phases, and these have been well reviewed (Robinson et al., 2016). Many methods for the analysis of microbiome datasets assume that sequencing data are equivalent to ecological data where the counts of reads assigned to organisms are often normalized to a constant area or volume. Methods applied include count-based strategies such as Bray-Curtis dissimilarity, zero-inflated Gaussian models and negative binomial models (McMurdie and Holmes, 2014; Weiss et al., 2017).

In an ecological study it is possible for many different species to co-exist, and their absolute abundance may be important. For example, in an area containing only tigers, it is important to know if the population size is sufficient to maintain needed genetic diversity for long-term survival (Shaffer, 1981). However, the abundance of one species may not influence the abundance of another; the area may contain both tigers and ladybugs, and the migration of several ladybugs into the area would not be expected to affect the number of tigers.

The assumption of true independence can not hold in high-throughput sequencing (HTS) experiments because the sequencing instruments can deliver reads only up to the capacity of the instrument. Thus, it is proper to think of these instruments as containing a fixed number of slots which must be filled. Returning to our tiger and ladybug analogy, the migration of ladybugs into an area containing a fixed number of slots that are already filled must displace either tigers or ladybugs from the occupied slots. This analogy extents, without restriction, to any number of taxa, and to any fixed capacity instrument (Aitchison, 1986; Lovell et al., 2011; Friedman and Alm, 2012; Fernandes et al., 2013, 2014; Lovell et al., 2015; Mandal et al., 2015; Gloor et al., 2016a,b; Gloor and Reid, 2016; Tsilimigras and Fodor, 2016). Thus, the total read count observed in a HTS run is a fixed-size, random sample of the relative abundance of the molecules in the underlying ecosystem. Moreover, the count can not be related to the absolute number of molecules in the input sample as shown in **Figure 1**. This is implicitly acknowledged when microbiome datasets are converted to relative abundance values, or normalized counts, or are rarefied (McMurdie and Holmes, 2014; Weiss et al., 2017) prior to analysis. Thus the number of reads obtained is irrelevant, and contains only information on the precision of the estimate (Fernandes et al., 2013). Data that are naturally described as proportions or probabilities, or with a constant or irrelevant sum, are referred to as compositional data. Compositional data contains information about the relationships between the parts (Aitchison, 1986; Pawlowsky-Glahn et al., 2015).

Data about a microbiome collected by high throughput sequencing are often examined under the assumption that sequencing is, in some way, *counting the number of molecules associated with the bacteria in the population*, as illustrated by the top barplot in **Figure 1B**. We can see the difference between counts and compositions by comparing the data for the actual counts for three samples in the top barplot with their proportions in the bottom barplot. Note, that samples 2 and 3 in **Figure 1B** have the same proportional abundances even though they have different absolute counts prior to sequencing. The difference in apparent direction of change is shown in **Figure 1C** and we can observe that the relationship between absolute abundance in the environment and the relative abundance after sequencing is not predictable.



	1→2	1→3	2→3
A	— ⇧	⇧ ⇧	⇧ —
B	⬇ ⬇	⬇ ⬇	⇧ —

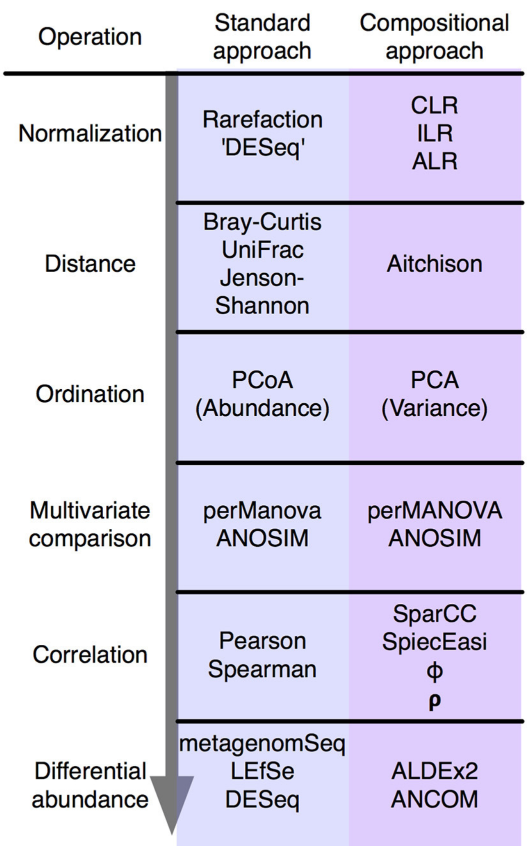
FIGURE 1 | High-throughput sequencing data are compositional. **(A)** illustrates that the data observed after sequencing a set of nucleic acids from a bacterial population cannot inform on the absolute abundance of molecules. The number of counts in a high throughput sequencing (HTS) dataset reflect the proportion of counts per feature (OTU, gene, etc.) per sample, multiplied by the sequencing depth. Therefore, only the relative abundances are available. The bar plots in **(B)** show the difference between the count of molecules and the proportion of molecules for two features, A (red) and B (gray) in three samples. The top bar graphs show the total counts for three samples, and the height of the color illustrates the total count of the feature. When the three samples are sequenced we lose the absolute count information and only have relative abundances, proportions, or "normalized counts" as shown in the bottom bar graph. Note that features A and B in samples 2 and 3 appear with the same relative abundances, even though the counts in the environment are different. The table below in **(C)** shows real and perceived changes for each sample if we transition from one sample to another.

2. PROBLEMS WITH CURRENT METHODS OF ANALYSIS

We will briefly outline the problems that arise when compositional data are examined using a non-compositional paradigm, stepping through the usual stages of analysis shown in **Figure 2**. All these issues have been extensively reviewed and debated in both the older and the more recent literature in fields as diverse as economics, geology and ecology. Thus, rather than present an exhaustive explanation of the problems, we will outline the major issue and cite a few useful resources.

It is very difficult to collect exactly the same number of sequence reads for each sample. This can be because of differences in platform (e.g., MiSeq vs. HiSeq) or because of technical difficulties in loading the same molar amounts of the sequencing libraries on the instrument, or because of random variation. The total number of counts observed (often referred to as read depth) is a major confounder for distance or dissimilarity calculations for multivariate ordinations derived from these distances (McMurdie and Holmes, 2014). Initial attempts in the microbiome field used "rarefaction" or subsampling of the read counts of each sample to a common read depth to attempt to correct this problem (Lozupone et al., 2011; Wong et al., 2016). The use of subsampling has been questioned since it results in a loss of information and precision (McMurdie and Holmes, 2014), and the practice of count normalization from the RNA-seq field has been advocated instead. There are a number of count normalization methods used and two, the trimmed mean of *M* values (TMM) (Robinson and Oshlack, 2010), and the median method (Anders and Huber, 2010) are similar to a log-ratio transformations, but are less suitable in highly asymmetrical or sparse datasets (Fernandes et al., 2013; Gloor et al., 2016a). These transformations are further undesirable since the number of counts observed by the instrument, by design, can not contain any information on the actual number of molecules in the environment, and because the investigator naturally interprets the results as counts instead of log-ratios.

One of the first analysis steps in a traditional analysis, following rarefaction or count normalization, is the calculation of a distance or dissimilarity (DD) matrix from the data that is used



Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

FIGURE 2 | The standard microbiome analysis tool kit and the compositional replacements. A simplified standard microbiome computational workflow is illustrated. The initial normalization steps are not formally equivalent since compositional data are inherently "normalized", and read count normalization is unnecessary. The other steps are functionally equivalent and substitute a compositionally appropriate approach for one that is not.

for downstream analyses such as ordination, and discrimination. Distances between features are non-linear when examined from a Euclidian perspective (Martín-Fernández et al., 1998; Aitchison et al., 2000) and many DD matrices are used that partially address this problem. As noted above the total number of reads in a sample is a strong confounding variable on all these methods, indicating that the composition of the sample is not the primary property being measured. However, apparently useful DD matrices can be generated after normalization. Three DD matrices dominate the literature; UniFrac (both the weighted and unweighted variants) (Lozupone et al., 2011), Bray-Curtis and Jensen-Shannon divergence, and while all have their uses, they do not account for the compositional nature of the data. It should be noted that the weighted UniFrac distance approach captures important phylogenetic information, and a recent compositional replacement has been developed (Silverman et al., 2017).

The major uses for the DD matrices are ordination and clustering. Here, the shortcomings of these DD methods become apparent. In addition to being sensitive to the total read depth of a sample, DD methods largely discriminate between samples based on the most relatively abundant features in the samples, not on the features that are necessarily the most variable between samples (Gorvitovskaia et al., 2016; Wong et al., 2016). This can lead to the location of samples in an ordination changing dramatically when different features are included or excluded from the dataset, and to a lack of sensitivity in identifying outlier samples (Wong et al., 2016).

Severe problems with correlation in compositional data were first noted at the dawn of statistical practice by Pearson (1897) and rediscovered in the context of microbiome studies (Lovell et al., 2011; Friedman and Alm, 2012; Lovell et al., 2015; Kurtz et al., 2015; Morton et al., 2017). Unfortunately, the effect cannot be diluted away as has been recommended (Weiss et al., 2016). Understanding that there is a correlation problem is crucial, since unconstrained correlation or covariation are key concepts for ordination, clustering, network analysis and differential (relative) abundance determination. Compositional data have a negative correlation bias and a different correlation structure than the underlying count data. Even worse, compositional data exhibit spurious correlation upon subsetting or aggregation. The "Correlation" section in the Supplement shows that correlation is not a reliable or a reproducible indicator of the underlying data when dealing with compositional data.

Finally, differential (relative) abundance measures do not account for compositionality (Fernandes et al., 2013; Mandal et al., 2015; Gloor et al., 2016a). Large scale tool benchmarking has revealed that differential (relative) abundance tools in common use are sensitive to sparsity (Thorsen et al., 2016) and consequently exhibit unacceptably high false positive identification rates (Hawinkel et al., 2017).

In summary the analysis of compositional data using current protocols has several challenges. However, as shown below these issues can be addressed in a satisfactory way using tools that account for the compositional nature of the data.

3. ANALYSIS OF HTS USING CODA METHODS

Compositional datasets from HTS can be analyzed in a rigorous manner by adapting tools from other fields (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015) and using new tools based on the same underlying foundations (Fernandes et al., 2013; Erb and Notredame, 2016; Silverman et al., 2017; Quinn et al., 2017). There are now examples in the literature that provide guidance on how to do some or all of these analyses on HTS datasets, including meta-transcriptomics (Macklaim et al., 2013) and tag-sequencing (McMurrough et al., 2014; Bian et al., 2017). We briefly review the approaches below.

The starting point for any compositional analyses is a ratio transformation of the data. Ratio transformations capture the relationships between the features in the dataset and these ratios are the same whether the data are counts or proportions. Taking the logarithm of these ratios, thus log-ratios, makes the data symmetric and linearly related, and places the data in a log-ratio coordinate space (Pawlowsky-Glahn et al., 2015). Thus, we can obtain information about the log-ratio abundances of features *relative to other features* in the dataset, and this information is directly relatable to the environment. We cannot get information about the absolute abundances since this information is lost during the sequencing process as explained in **Figure 1**. However, log-ratios have the nice mathematical property that their sample space is real numbers, and this represents a major advantage for the application of standard statistical methods that have been developed for real random variables.

Often the centered log-ratio (clr) transformation introduced by Aitchison (1986) is used. Given an observation vector of D "counted" features (taxa, operational taxonomic units or OTUs, genes, etc.) in a sample, $\boldsymbol{x} = [x_1, x_2, ...x_D]$, the clr transformation for the sample can be obtained as follows:

$$\begin{aligned} \boldsymbol{x}_{clr} &= [log(x_1/G(\boldsymbol{x})), log(x_2/G(\boldsymbol{x})) \ldots log(x_D/G(\boldsymbol{x}))], \\ G(\boldsymbol{x}) &= \sqrt[D]{x_1 \cdot x_2 \cdot ... \cdot x_D} \end{aligned} \quad (1)$$

$G(\boldsymbol{x})$ is the geometric mean of $\boldsymbol{x}$. The clr transformed values can be used as inputs for multivariate hypothesis testing using tools such as MANOVA, regression etc. (Van den Boogaart and Tolosana-Delgado, 2013) and for model building. The clr-transformed values are scale-invariant; that is the same ratio is expected to be obtained in a sample with few read counts or an identical sample with many read counts, only the precision of the clr estimate is affected. This is elaborated in the "Probability" and "Log-ratio transformations" section in the Supplement, but the consequence is that count normalization is unnecessary and indeed, undesirable since information on precision is lost.

The $G(\boldsymbol{x})$ cannot be determined for sparse data without deleting, replacing or estimating the 0 count values. Fortunately, there are acceptable methods of dealing with 0 count values as both point estimates using `zCompositions R` package (Palarea-Albaladejo and Martín-Fernández, 2015), and as a probability distribution using `ALDEx2` available on Bioconductor. Converting the single estimate to a probability vector prior to clr transformation produces a scale-invariant measure since this accounts for the precision of the estimate of the probabilities for each feature; we refer advanced readers to the more technical literature (Jaynes and Bretthorst, 2003; Fernandes et al., 2013; Gloor et al., 2016a) and the "Probability" section of the Supplement for more information.

There are compositional replacements for distance determination that is used for clustering and ordination. The first is the `philr` phylogenetic transform (and `R` package) based on balances (binary partitions) along an evolutionary tree (Silverman et al., 2017) that is a replacement for the familiar UniFrac distance metric. Distances determined by phylogenetic transforms have the advantage that the binary partitions chosen have a simple interpretation and the correlation structure of the data is fully accounted for. However, the disadvantage is that only the relationships between the chosen partitions can be examined. A second distance metric is the Aitchison distance, which is simply the Euclian distance between samples after clr transformation, and the distances between samples are the same as the phylogenetic ilr. The Aitchison distance is superior to both the widely used Jensen-Shannon divergence and the Bray-Curtis dissimilarity metrics, being more stable to subsetting and aggregating of the data, and being a true linear distance (Aitchison et al., 2000).

The replacement for β-diversity exploration of microbiome data is the variance-based compositional principal component (PCA) biplot (Aitchison, 1983; Aitchison and Greenacre, 2002) where the relationship between inter-OTU variance and sample distance can be observed (Gloor et al., 2016b). The compositional biplot has several advantages over the principal co-ordinate (PCoA) plots for β-diversity analysis. The results obtained are very stable when the data are subset (Bian et al., 2017), meaning that exploratory analysis is not driven simply by the presence absence relationships in the data nor by excessive sparsity (Wong et al., 2016; Morton et al., 2017). PCA plots can be substantially more reproducible, since they do not depend upon an presumed underlying tree that may need to be regenerated with each data subset, or when new taxa need to be incorporated. This simplicity facilitates exploratory data analysis. Compositional PCA biplots display the relationships between OTUs and the distances between samples on a common plot. It is possible to glean substantial qualitative information regarding the quality of the dataset and the relationships between groups with this tool (Aitchison and Greenacre, 2002; Gloor et al., 2016b), and examples are shown in the "Biplot" section of the Supplement.

As noted above, the correlation is unreliable in compositional datasets because of the negative correlation bias and the instability of correlation to subsetting the data. This is explained more fully in the supplement (Pearson, 1897; Aitchison, 1986) but these problems are observed with all non-compositional correlation methods (Ortego and Egozcue, 2013). Unfortunately, correlation cannot be subjected to a principled process to determine the optimal method as has been advocated recently (Weiss et al., 2016).

There are several more rigorous approaches that can be applied to analyze correlation in microbiome datasets, including SPARCC (Friedman and Alm, 2012) and SPieCeasi (Kurtz et al., 2015), both of which assume a sparse data matrix, and the ϕ (Lovell et al., 2015) and ρ (Erb and Notredame, 2016) metrics (the published versions of which required a non-sparse matrix). These latter two metrics have been incorporated into the `R` package `propr`, that includes an adaptation allowing the calculation of the metrics with sparse data that gives an expected value of ρ (E(ρ)), that approaches 1 if the two features have exactly constant ratios in the data (Lovell et al., 2015; Quinn et al., 2017). Supplementary Figure 2 shows that the expected value of ρ is much more stable to subsetting than are familiar correlation metrics, and becomes more reproducible as the value of E(ρ) approaches 1, thus indicating greater precision in estimation as correlation becomes stronger. However, determining an optimal and general approach for correlation in compositional datasets is an open research problem. Supplementary Figures 2–5 have a more extended explanation of the correlation problem and the use of E(ρ) as a proposed solution.

Differential (relative) abundance of OTUs between groups in compositional data is often examined using purpose-built tools that compare the difference in relative abundance across samples, and recently tools adapted from the domain of RNA-seq have been suggested. Unfortunately, these approaches do not account for the compositional nature of the data, and so can be particularly sensitive to the negative correlation bias and large variability of such datasets (Fernandes et al., 2013). Indeed benchmarking suggests that traditional tools exhibit different false positive rates with different levels of sparsity (Thorsen et al., 2016), and that the false positive rates can be up to 20× higher than expected (Hawinkel et al., 2017).

Tools based on an approximate compositional foundation are available. The ANCOM tool performs statistical tests on point estimates of data transformed by an additive log ratio, where (presumed) invariant taxa are chosen as the denominator (Mandal et al., 2015). ANCOM is being incorporated into the popular QIIME suite of microbiome analysis tools (Weiss et al., 2017). The ALDEx2 tool performs statistical tests on the clr values from a modelled probability distribution of the dataset (Supplementary data Equations 1–4), and reports the expected values of parametric and non-parametric statistical tests along with effect-size estimates. This approach reduces the false-positive identification problem to near 0 in real and modelled microbiome datasets with little effect of sensitivity (Thorsen et al., 2016) and is observed to be relatively insensitive to change when the data are subset (Fernandes et al., 2014). There are many examples in the literature on its use (Macklaim et al., 2013; McMurrough et al., 2014; Bian et al., 2017) and in the Supplementary.

In summary, the analysis of compositional data by traditional methods can appear to give satisfactory results. However, these results can be misleading and unpredictable. Compositionally-appropriate tools exist as drop-in replacements at each stage of the analysis as shown in **Figure 2**, and interested readers are directed to the supplementary and to other published examples (Macklaim et al., 2013; Fernandes et al., 2014; McMurrough et al., 2014; Lovell et al., 2015; Mandal et al., 2015; McMillan et al., 2015; Gloor and Reid, 2016; Gloor et al., 2016b; Bian et al., 2017; Silverman et al., 2017; Quinn et al., 2017), and the similar correspondence analysis implemented in the phyloseq package (McMurdie and Holmes, 2013).

AUTHOR CONTRIBUTIONS

GG conceived and wrote the initial draft of the manuscript. JM conceived and made **Figures 1**, **2**. JM, JE, and VP-G edited the draft. All authors agreed with the contents of the final version.

FUNDING

Work by GG and JM were supported by grants from the National Sciences and Engineering Research Council grant RGPIN-03878-2015, and by funding from the Ministry of Agriculture and Agri-Food AgriInnovation Program. VP-G and JE acknowledge financial support by the Spanish Ministry of Education and Science under project 'CODA-RETOS' (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full#supplementary-material

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65. doi: 10.1093/biomet/70.1.57

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Math. Geol.* 32, 271–275. doi: 10.1023/A:1007529726302

Aitchison, J., and Greenacre, M. (2002). Biplots of compositional data. *J. Roy. Stat. Soc. Ser. C* 51, 375–392. doi: 10.1111/1467-9876.00275

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

Bian, G., Gloor, G. B., Gong, A., Jia, C., Zhang, W., Hu, J., et al. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere* 2:e00327-17. doi: 10.1128/mSphere.00327-17

Erb, I., and Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory Biosci.* 135, 21–36. doi: 10.1007/s12064-015-0220-8

Fernandes, A. D., Macklaim, J. M., Linn, T., Reid, G., and Gloor, G. B. (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE* 8:e67019.

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, 15.1–15.13. doi: 10.1186/2049-2618-2-15

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

Gloor, G. B., Macklaim, J. M., Vu, M., and Fernandes, A. D. (2016a). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Aust. J. Stat.* 45, 73–87. doi: 10.17713/ajs.v45i4.122

Gloor, G. B., and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* 62, 692–703. doi: 10.1139/cjm-2015-0821

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016b). It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* 26, 322–329. doi: 10.1016/j.annepidem.2016.03.003

Gorvitovskaia, A., Holmes, S. P., and Huse, S. M. (2016). Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome* 4:15. doi: 10.1186/s40168-016-0160-7

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. Bioinf.* bbx104. doi: 10.1093/bib/bbx104

Jaynes, E. T., and Bretthorst, G. L. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press

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

Lovell, D., Müller, W., Taylor, J., Zwart, A., and Helliwell, C. (2011). "Proportions, percentages, ppm: do the molecular biosciences treat compositional data right," in *Compositional Data Analysis: Theory and Applications,* eds V. Pawlowsky-Glahn and A. Buccianti (London: John Wiley & Sons, Ltd.), 193–207.

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

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). Unifrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Macklaim, M. J., Fernandes, D. A., Di Bella, M. J., Hammond, J.-A., Reid, G., and Gloor, G. B. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* 1:15. doi: doi: 10.1186/2049-2618-1-12

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

Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., and Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. *Proc. IAMG.* 98, 526–531.

McMillan, A., Rulisa, S., Sumarah, M., Macklaim, J. M., Renaud, J., Bisanz, J. E., et al. (2015). A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. *Sci. Rep.* 5, 14174. doi: 10.1038/srep14174

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8:e61217. doi: 10.1371/journal.pone.0061217

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

McMurrough, T. A., Dickson, R. J., Thibert, S. M. F., Gloor, G. B., and Edgell, D. R. (2014). Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2376–E2383. doi: 10.1073/pnas.1322352111

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., and Knight, R. (2017). Uncovering the horseshoe effect in microbial analyses. *mSystems* 2:e00166-16. doi: 10.1128/mSystems.00166-16

Ortego, M. I., and Egozcue, J. J. (2013). "Spurious copulas," in *Proceedings of the 5th Workshop on Compositional Data Analysis, CoDaWork 2013* (Vorau).

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intel. Lab. Syst.* 143, 85–96. doi: 10.1016/j.chemolab.2015.02.019

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.* London: John Wiley & Sons.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. – on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. Roy. Soc. Lond.* 60, 489–498.

Quinn, T., Richardson, M. F., Lovell, D., and Crowley, T. (2017). propr: An R-package for identifying proportionally abundant features using compositional data analysis. *bioRxiv*. doi: 10.1101/104935

Robinson, C. K., Brotman, R. M., and Ravel, J. (2016). Intricacies of assessing the human microbiome in epidemiologic studies. *Ann. Epidemiol.* 26, 311–321. doi: 10.1016/j.annepidem.2016.04.005

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25.1–R25.9. doi: 10.1186/gb-2010-11-3-r25

Shaffer, M. L. (1981). Minimum population sizes for species conservation. *BioScience* 31, 131–134.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* 6:21887. doi: 10.7554/eLife.21887

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

Tsilimigras, M. C. B., and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330–335. doi: 10.1016/j.annepidem.2016.03.002

Van den Boogaart, K. G., and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*, London, UK: Springer.

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

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

Wong, R. G., Wu, J. R., and Gloor, G. B. (2016). Expanding the UniFrac toolbox. *PLoS ONE* 11:e0161196. doi: 10.1371/journal.pone.0161196

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 Microbiology

METHODS
published: 07 November 2017
doi: 10.3389/fmicb.2017.02114

Analysis of Microbiome Data in the Presence of Excess Zeros

Abhishek Kaul [1†], *Siddhartha Mandal* [2], *Ori Davidov* [3] *and Shyamal D. Peddada* [1*†]

[1] *Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIH), Durham, NC, United States,* [2] *Public Health Foundation of India, Gurgaon, India,* [3] *Department of Statistics, University of Haifa, Haifa, Israel*

Motivation: An important feature of microbiome count data is the presence of a large number of zeros. A common strategy to handle these excess zeros is to add a small number called pseudo-count (e.g., 1). Other strategies include using various probability models to model the excess zero counts. Although adding a pseudo-count is simple and widely used, as demonstrated in this paper, it is not ideal. On the other hand, methods that model excess zeros using a probability model often make an implicit assumption that all zeros can be explained by a common probability models. As described in this article, this is not always recommended as there are potentially three types/sources of zeros in a microbiome data. The purpose of this paper is to develop a simple methodology to identify and accomodate three different types of zeros and to test hypotheses regarding the relative abundance of taxa in two or more experimental groups. Another major contribution of this paper is to perform constrained (directional or ordered) inference when there are more than two ordered experimental groups (e.g., subjects ordered by diet or age groups or environmental exposure groups). As far as we know this is the first paper that addresses such problems in the analysis of microbiome data.

Results: Using extensive simulation studies, we demonstrate that the proposed methodology not only controls the false discovery rate at a desired level of significance while competing well in terms of power with DESeq2, a popular procedure derived from RNASeq literature. As expected, the method using pseudo-counts tends to be very conservative and the classical t-test that ignores the underlying simplex structure in the data has an inflated FDR.

Keywords: Microbiome data, Aitchisons log-ratio, bootstrap, covariates, cross-sectional data, false discovery rate (FDR)

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Magnus Øverlie Arntzen,
Norwegian University of Life Sciences,
Norway
Bradley Stevenson,
University of Oklahoma, United States

***Correspondence:**
Shyamal D. Peddada
sdp47@pitt.edu

†*Present Address:*
Abhishek Kaul,
Department of Mathematics and Statistics, Washington State University, Pullman, WA, United States
Shyamal D. Peddada,
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, United States

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

Received: *25 May 2017*
Accepted: *17 October 2017*
Published: *07 November 2017*

Citation:
Kaul A, Mandal S, Davidov O and Peddada SD (2017) Analysis of Microbiome Data in the Presence of Excess Zeros. Front. Microbiol. 8:2114. doi: 10.3389/fmicb.2017.02114

1. INTRODUCTION

Microbial count data are represented using operational taxonomic units (OTUs) from 16S rRNA studies. For each specimen (e.g. fecal sample) drawn from an ecosystem (e.g. gut), the number of occurrences of each OTU is measured and the resulting OTU table is summarized to obtain relative abundance for bacterial taxa in a specimen. These OTU counts may be summarized at any level of the bacterial phylogeny, e.g., species, genus, family, order, etc. Throughout this paper we use the generic term “taxa” to denote a particular phylogenetic classification. Since the relative abundances of taxa in a specimen sum to 1, these are compositional data and they reside in a simplex rather than the entire Euclidean space. Another important feature of these microbiome data is that not all taxa may be present in each sample, i.e., some of the OTUs may take zero values. Using such microbial compositional data, researchers are interested in understanding the interplay between microbiome, diet, genome and human health (Clemente et al., 2012; den Besten et al., 2013). Accordingly, there is an urgent need for statistical methods for analyzing these complex microbial count data. This is an active area of research and a variety of statistical and computational methods have been proposed in the literature to answer a variety of scientific questions. For a review one may refer to Li (2015) and Mandal et al. (2015). The latter described in detail various statistical parameters associated with microbial compositional data and discuss which are estimable, and hence testable, and which are not. They proposed Aitchison's log-ratio based methodology (Aitchison, 1982, 1985, 1986) called ANCOM for comparing the taxa abundance at the ecosystem level in two or more groups or populations. Earlier, Xia et al. (2013) also considered Aitchison's log-ratio based methodology for microbiome data and proposed a penalized likelihood based methodology to select covariates influencing microbiome expression.

Excess zeros in microbiome data present a challenge when analyzing these data, specifically when comparing two or more experimental groups. A common strategy to handle these excess zeros is to add a small number called pseudo-count (e.g., 1, cf. Xia et al., 2013; Mandal et al., 2015). Although adding a pseudo-count appears to be a reasonable and a simple strategy, it is *ad-hoc*. Other strategies include modeling excess zeros using various probability models (Paulson et al., 2013; Chen and Li, 2016). However, such models often make an implicit assumption that all zeros can be explained by a common probability model. As described in this article, this is not always the case as there are potentially three different sources of zeros in microbiome data. The first major contribution of this paper is a method which identifies the three major types or sources of zeros in microbiome data. The second major contribution of this paper is to compare the mean relative abundance of taxa in two or more groups while taking into consideration the compositional structure and the type of zeros in the data. Unlike ANCOM (Mandal et al., 2015), which compares the taxa abundance in the ecosystem of two or more groups, the proposed methodology compares the abundance of taxa relative to a background value. The method is general enough that the reference background value can be a specific taxon the user is interested in or it can be some suitable background value specific to each specimen, such as the geometric mean (Aitchison's centered log-ratios). The main idea is to normalize data within each specimen so that any background values within the specimen are eliminated. This idea is analogous to what is often done in gene expression studies. If a particular taxon is used as the reference taxon or reference value , then we assume that the taxon is present in all specimens. Thus the normalizing variable is same across all specimens. From our experience, in practice this condition is not particularly stringent, especially if the researcher is interested in studying microbiome at the genus or a higher level of the phylogenetic tree. For example, in the Yatsunenko et al. (2012) study consisting of 531 samples over three geographical locations (US, Venezuela and Malawi) there exist at least one taxon (at the genus level) that is present in all samples. These data are discussed later in this manuscript. If no such taxon exists, then the proposed methodology can be implemented using the geometric mean as the reference to correct for the background abundance levels of each specimen.

In some applications researchers are interested in performing inferences regarding mean relative abundances of individual taxon in the ecosystems of more than two ordered groups. For example, one may be interested in comparing the mean relative abundances of individual taxon in subjects ordered by different levels of fat intake or levels of dietary supplements or subjects belong to different age groups etc. In all such situations the classical two-sided tests are not as informative or powerful as the constrained inference (or order restrictions) based tests (Farnan et al., 2014; Jelsema and Peddada, 2016). Since the proposed methodology converts the simplex data to Euclidean space data, constrained inference theory developed in Farnan et al. (2014) is directly applicable to the present setting. Thus the third major contribution of this paper is to perform constrained inference when there are more than two ordered experimental groups. As far as we know this is the first paper that addresses such problems in the analysis of microbiome data. Owing to the generality of Farnan et. al. methodology to (a) cross-sectional as well as repeated measures/longitudinal designs, (b) detecting trends in the relative abundances of taxa in two or more ordered experimental groups such as in time course experiments, dose-response studies or when comparing subjects at stages of disease, (c) multiple pairwise comparisons of several experimental groups against a pre-specified control group, the methodology described in this paper is therefore very broadly applicable. Thus, the proposed methodology can be used for testing a wide range of hypotheses while controlling for false discovery rate (FDR) at the desired nominal level. Extensive simulations are performed to demonstrate that the proposed methodology controls the FDR in a variety settings considered in the simulation study while enjoying higher power than some commonly used methods including those based on pseudo-counts. We illustrate the methodology using the global gut data of Yatsunenko et al. (2012).

2. NOTATION AND PROBLEM FORMULATION

Suppose a sample of n_j specimens are drawn from the j^{th} experimental group, $j = 1, 2, \ldots, J$. On each specimen suppose the abundance of p taxa are obtained. Here the word "taxa" could be at any level of the bacterial phylogeny, e.g., species, genus, family, order, etc., or just the counts of OTU categories themselves. Let z_{ijk} denote the observed abundance of k^{th} taxon, $k = 1, 2, \ldots, p$, in the i^{th} specimen from the j^{th} experimental group. In vector notation we have $z_{ij} = (z_{ij1}, \ldots, z_{ijp})$. For simplicity of exposition throughout this paper, we shall take $n_j = n$, $j = 1, 2, \ldots, J$ even though the methodology does not require the design to be balanced. As explained in Mandal et al. (2015), unlike most commonly encountered biological data, the basic counts of OTU categories within each specimen cannot be regarded as absolute values but only relative values as they depend upon the sampling depth corresponding to each specimen. In other words, it does not make sense to compare the expected value of the observed counts between two experimental groups. To draw any meaningful inferences regarding the taxa abundance in two or more groups one needs to "normalize" the data within each specimen. Since classical inference, such as t-tests or ANOVA are not valid in the present context due to the simplex constraint, following Aitchison (1980) and Mandal et al. (2015) worked with log-ratios of relative abundances within each specimen. This is equivalent to computing log-ratios of abundances of each taxon relative to a "reference value." Thus, for the i^{th} specimen in the j^{th} experimental group, one may consider the following expression to normalize the data z_{ijk}:

$$\log z_{ijk} - f_{ij}(z_{ij1}, \ldots, z_{ijp}), \tag{2.1}$$

using some pre-specified "reference value" $f_{ij}(z_{ij1}, \ldots, z_{ijp})$. For example, $f_{ij}(z_{ij1}, ..., z_{ijp}) = \log z_{ijb}$, where z_{ijb} is the count corresponding to a pre-specified reference taxon b. Alternatively, using the non-zero values z_{ijk}, $k = 1, 2, ..., p$, the user may choose $f_{ij}(z_{ij1}, \ldots, z_{ijp}) = r^{-1} \sum_{\{k\,:\,z_{ijk} \neq 0\}} \log z_{ijk}$, where r is the number of non-zero components in $(z_{ij1}, z_{ij2}, \ldots, z_{ijp})'$, i.e., the logarithm of the geometric mean of the OTU counts within each experimental group $j = 1, \ldots, J$ (Aittchisons centered log-ratio).

Although the above normalization procedure eliminates the effect of the library size within specimen, it does not account for differences in the library sizes across specimens. To deal with this, we make another correction to the above normalization step. We make the assumption that all specimens within an experimental group are a random sample from a common population of specimens so that the observed background value for a given specimen is a random realization from a common population of all background values. Thus we have the following one-way ANOVA model describing the observed background value:

$$f_{ij}(z_{ij1}, \ldots, z_{ijp}) = \mu_j + \varepsilon_{ij}, \tag{2.2}$$

where μ_j is the fixed effect due to the experimental group $j = 1, \ldots, J$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon)$ is a random variable that captures variation due to the sampling depth. This quantity can then be predicted by the residual $\hat{\varepsilon}_{ij} = f_{ij}(z_{ij1}, \ldots, z_{ijp}) - \frac{1}{n} \sum_{i \in j^{th} \text{ group}} f_{ij}(z_{ij1}, \ldots, z_{ijp})$ which can be interpreted as the best linear unbiased predictor (BLUP) in the assumed model.

Hence in place of the typical normalization (2.1), we normalize the raw abundances using the following normalized formula:

$$y_{ijk} = \log z_{ijk} - \Big(f_{ij}(z_{ij1}, ..., z_{ijp}) - \hat{\mu}_j\Big) \tag{2.3}$$

where $\hat{\mu}_j = \frac{1}{n} \sum_{i \in j^{th}\ group} f_{ij}(z_{ij1}, ..., z_{ijp})$. This normalization procedure can be easily extended to the case when there are covariates present in the model. Of course, in the above formula, all logarithms are calculated under the assumption that there are no zero values. However, as mentioned earlier, this is not true with the microbiome data. We address this problem in the next section.

3. ZEROS

A special feature of a microbiome data matrix is that it is higly sparse, i.e., a very high proportion of data entries are zero (absent taxa). For example, at the genera level, nearly 80% of the data matrix in the Global gut data of Yatsunenko et al. (2012) are zero. Furthermore, corresponding to a given taxon, the counts may vary from 0 to the order of 10^5 across samples within an experimental group. In this section we develop a pre-processing step that not only helps us potentially understand the different types of zeros in the data but address them accordingly.

3.1. Outlier Zeros

For a given taxon k in the j^{th} group, we declare the sample i to be an "outlier zero" if its count is zero and is declared to be an outlier by the methodology described below. In our assessment, this taxon is recorded as zero due to some extraneous reasons but not because it is below detection limits due to sampling depth. Thus, as far as taxon k is concerned, the i^{th} sample within group j is an outlier.

We first convert the count data into continuous scale by adding a pseudo-count of 1 and normalize the data using the transformation pseudo-count (2.3). Let $y_{ij} = (y_{ij1}, ..., y_{ijp})$ denote the p dimensional vector for i^{th} observation in the j^{th} group, then for each j, k, we model y_{ijk} using the following mixture of normal distributions. Since our outlier detection algorithm is applied to each experimental group j and each taxon k, for simplicity of exposition, we drop the subscript j and k from the following:

$$y_i \sim^{i.i.d} \pi \mathcal{N}(\mu_1, \sigma_1) + (1 - \pi)\mathcal{N}(\mu_2, \sigma_2), \;\; i = 1, \ldots, n \tag{3.1}$$

The main idea of our methodology is that when means of the two normal distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$ in the above mixture are "well separated and the left cluster, i.e. cluster corresponding to mean μ_1, forms only a small fraction of the total number of observations of the group, i.e. π is small, then it is reasonable to assume that the left cluster is a collection of outlier observations in the group and the observed zero might be a potential outlier. On the other hand, if the two groups are not well separated then the observed zero may not be an outlier zero but zero due to other reasons. Such zeros are handled later in this section.

Identification of two clusters: For a given taxon within a group, we declare that its distribution is a mixture of two "distant" normal distributions if the following two criteria are satisfied:

1. **Separation**: The 97.5th percentile of the first distribution does not overlap with the 2.5th percentile of the second distribution, i.e., $\mu_1 + 1.96\sigma_1 < \mu_2 - 1.96\sigma_2$.
2. **Frequency**: One distribution is "c % heavier" than other, i.e., $\pi < c$ for some pre-specified c.

The above determinations, along with the estimation of parameters $\pi, \mu_1, \mu_2, \sigma_1, \sigma_2$ of the mixture (3.1) can be performed efficiently by an algorithm due to Peddada and Hwang (2002). We refer to the data cells identified by this mechanism as "outlier zeros" which are ignorable entries (replaced by NA in the data).

3.2. Structural Zeros

In many cases, because of the nature of the experimental groups, some taxa are not supposed to be present in samples obtained from some groups but may be present in others. For example, babies exposed to antibiotics may be devoid of some taxa in their fecal samples, which are present in healthy babies not exposed to antibiotics. Although, in theory the antibiotics exposed babies are expected to be completely devoid to some taxa, due to variability in the exposure and other factors, such taxa may not be 100% missing in the antibiotics exposed babies. Suppose p represents the proportion of non-zero taxa across all specimens in an experimental group. Then we expect p to be close to zero, if not exactly zero, in experimental groups where the taxon is not expected to be present. We refer to such zeros as structural zeros. For the j^{th} taxon in the k^{th} experimental group, let $\hat{p}_{jk} = \sum_{i=1}^{n} z_{ijk}\big/n$. Then we declare the taxon to have a structural zero value if either of the following is true.

1. $\hat{p}_{jk} = 0$
2. $\hat{p}_{jk} - 1.96\sqrt{\hat{p}_{jk}(1-\hat{p}_{jk})/n} \leq 0.$

Taxa that are identified as structural zeros in any given group are ignored from all future analyses for that group. Thus, for example, if in a study there are three experimental groups and if a particular taxon t is declared to have structural zero in Group 1 but not in Groups 2 and 3, then we automatically declare that taxon t is differentially abundant in Group 2 relative to Group 1 as well as in Group 3 relative to Group 1. We then compare the relative abundance of t between Groups 2 and 3 using the methodology developed in this paper.

3.3. Sampling Zeros

If an observed zero in the data does not qualify as an outlier zero or as a structural zero, then we declare such a zero to be sampling zero, perhaps caused by the sampling depth. In other words, these zeros are potentially due to the fact the taxon is relatively a rare taxon compared to other taxa in the specimen and due to technological (or other) reasons it was not observed. These sampling zeros are imputed by using a small pseudo-count value (e.g., 1) before analyzing the data. More generally, an imputation approach could also be applied to these left over zeros, however this is outside the scope of this manuscript.

To summarize, using the above process, we obtain a modified data set where; (a) samples with structural zeros are suitably removed from the data matrix, (b) the outlier zeros are treated as missing at random (MAR) in the sense of Rubin (1976) and the corresponding entries are replaced as “NA”, and (c) the sampling zeros are imputed as 1.

4. ANALYSIS OF TWO OR MORE GROUPS

In rest of this paper, we work with normalized data y described in Equation (2.3) after suitably dealing with zeros as described in the previous section. For the k^{th} taxon in the j^{th} experimental group, for $i = 1, 2 \ldots, n$, let $\mu_{jk} = E(y_{ijk)}$ and $\sigma_{jk}^2 = Var(y_{ijk})$. Using the zeros corrected data, we obtain the following unconstrained estimators for μ_{jk} and σ_{jk}^2, for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, p$:

$$\hat{\mu}_{jk} = \frac{\sum_{i=1}^{n} \mathbf{1}[y_{ijk} \neq NA] y_{ijk}}{\sum_{i=1}^{n} \mathbf{1}[y_{ijk} \neq NA]},$$
$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n} \mathbf{1}[y_{ijk} \neq NA](y_{ijk} - \hat{\mu})^2}{\sum_{i=1}^{n} \mathbf{1}[y_{ijk} \neq NA] - 1} \qquad (4.1)$$

In many applications, researchers are interested in comparing taxa relative abundances in two or more experimental groups. Depending upon the scientific question, one may perform a wide range of analyses. In this section we describe four different classes of analyses one may perform. In each case the statistical parameters of interest are μ_{jk}, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, p$. Note that, by construction, within each group j, $\sum_{k=1}^{p} \mu_{jk} = 0$. Hence without loss of generality, we limit rest of the discussion to the first $p-1$ taxa because $\mu_{jp} = -\sum_{k=1}^{p-1} \mu_{jk}$.

4.1. H_1: Two-Sided Global Hypotheses

Since the data y_{ijk} belong to the Euclidean space, therefore for each taxon k, $k = 1, 2, \ldots, p-1$, we can use standard linear model based methodology to test such hypotheses on the group means $\mu_{1k}, \mu_{2k}, \ldots, \mu_{Gk}$. adjusting for any covariates present in the data. If there are repeated measures or longitudinal data, then one can invoke the standard linear mixed effects models theory and test two-sided global hypotheses such as:

$$H_0 : \mu_{1k} = \mu_{2k} = \ldots = \mu_{Jk}$$

Vs.

$$\mu_{rk} \neq \mu_{sk},$$

for some $r \neq s$. The p-values obtained for each taxon k, $k = 1, 2, \ldots, p-1$, can be corrected for multiple testing using a suitable multiple testing correction procedure, such as Bonferroni or Benjamini-Hochberg (BH), depending upon the criterion of interest, namely, the Familywise error rate (FWER) or the false discovery rate (FDR).

4.2. H_2: Directional Multiple Pairwise Testing

For each taxon k, $k = 1, 2, \ldots, p-1$, often researchers are not interested in testing the global hypotheses H_1 but are interested in pairwise comparisons among some (or all) pre-specified experimental groups. Furthermore, within each pairwise comparison, a researcher may be interested in knowing if the (relative) abundance of a taxon increased or decreased from one group to the other. For example, a researcher may be interested in testing whether there is a greater (relative) abundance of *Bifidobacterium Sp.* in vaginally born babies who were never exposed to antibiotics during the first four months of life, than vaginally born babies who received at least one dose of antibiotics during the first four months. To draw such directional inferences in pairwise comparisons while controlling for the overall false discovery rate, one may apply the mdFDR (mixed directional FDR) controlling procedure introduced in Guo et al. (2010). When there are no covariates present, the Guo et al. (2010) procedure is available in the software

ORIOGEN 4.1. https://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/.

4.3. H_3: Directional Multiple Pairwise Testing against a Specific Experimental Group

Hypotheses H_2 deals pairwise comparisons among some pre-specified subset (or all) experimental groups. However, there are instances where researchers may be interested in testing all experimental groups against one pre-specified experimental group, such as, for example the control group. In such cases the power of Guo et al. (2010) procedure can be improved by appealing to the Dunnett's type test derived in Grandhi et al. (2016). The R-code for the method is provided in Grandhi et al. (2016).

4.4. H_4: Testing for Patterns

In some applications, a researcher may not be specifically interested in pairwise comparisons, but may be interested in detecting overall trends/patterns in the relative abundance of a taxon over multiple ordered (or partially ordered) experimental groups. Order (or partial order) among experimental groups arises when the experimental groups represent time or dose or stages of disease etc.

For example, researchers may be interested in understanding the trends in (relative) abundance of taxa across four partially ordered groups, namely, (G1) Vaginally born babies who were not exposed to any antibiotics during the first four months after birth, (G2) Vaginally born babies who were exposed to at least one dose of antibiotics during the first four months of after birth, (G3) C-Section born babies who were not exposed to any antibiotics during the first four months after birth and (G4) C-Section born babies who were exposed to at least one dose of antibiotics during the first four months of after birth. In this case, groups G1 and G4 are the extreme groups in terms of gut microbial environment. In G1 there are no interventions, and in G4 there are two interventions (C-section and antibiotics exposure). Groups G2 and G3 are intermediate groups with one intervention each (either C-Section or antibiotics exposure). Although, groups G2 and G3 are intermediate to G1 and G4, the order between G2 and G3 is uncertain and hence we have a partial ordering among the four groups.

A study design such as the one in this example can be represented using the **Figure 1C**, called a simple loop order, where, for each taxon, the researcher is interested in obtaining two sets of patterns, namely, pattern over G1, G2, and G4 and a pattern over G1, G3, and G4. Note that members within each set are completely ordered in terms of baby's exposure to interventions. When groups are ordered, one may be interested in identifying taxa whose mean relative abundance increases (or decreases) as we go from one extreme group (e.g., Group 1) to the other extreme group (e.g., Group 4) within each set. Such monotonic patterns, increasing or decreasing, are called the simple order (**Figure 1A**). More, precisely, for each taxon, $k = 1, 2, \ldots, p - 1$, one may be interested in testing the following hypotheses:

$$H_{10} : \mu_{1k} = \mu_{2k} = \mu_{4k}$$

Vs.

$$H_{1a} : \{\mu_{1k} \leq \mu_{2k} \leq \mu_{4k}\} \bigcup \{\mu_{1k} \geq \mu_{2k} \geq \mu_{4k}\},$$

and

$$H_{20} : \mu_{1k} = \mu_{3k} = \mu_{4k}$$

Vs.

$$H_{2a} : \{\mu_{1k} \leq \mu_{3k} \leq \mu_{4k}\} \bigcup \{\mu_{1k} \geq \mu_{3k} \geq \mu_{4k}\}.$$

In some applications one may be interested in identifying taxa that have an umbrella shaped pattern as in **Figure 1B**.

As observed above, rather than using some arbitrary parametric functions, one can describe various patterns or trends using mathematical inequalities, called order restrictions. To determine the best pattern or trend for each taxon we adopt the strategy in Peddada et al. (2003), where a similar problem was considered for time-course gene expression data. For each taxon, we test the null hypothesis that there is no change in mean relative abundance (in log scale) over all the experimental groups against the alternative hypothesis which is the union of all patterns of interest. For each pattern we construct a suitable order restricted test and the final test statistic is taken to be the maximum of all test statistics. The null distribution of the test statistic is derived using the residual bootstrap based procedure developed in Farnan et al. (2014) which is implemented in the package called constrained linear mixed effects (CLME), an R code developed by Casey Jelsema and is described in Jelsema and Peddada (2016). The R code allows for modeling covariates as well as longitudinal/repeated measurements data. Since there are

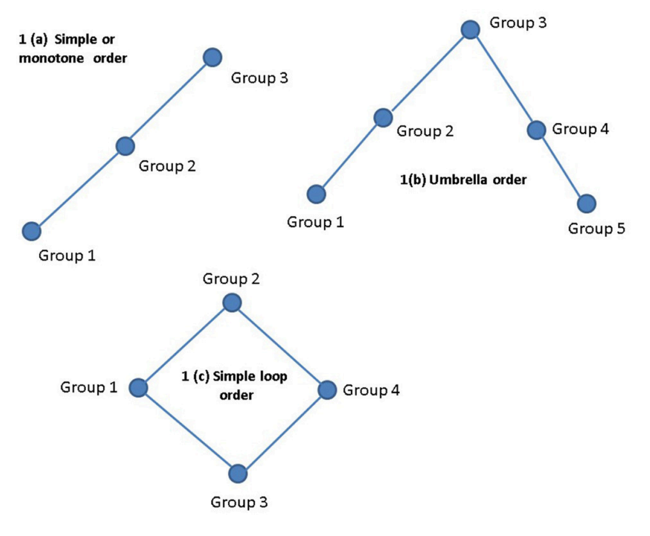


FIGURE 1 | Illustration of hypotheses H_{1a} and H_{2a} testing for trends amongst groups.

a large number of taxa, we perform multiple testing corrections using the BH procedure to control for the overall FDR. As in Peddada et al. (2003), if for a taxon, the null hypothesis is rejected at the desired level of significance (FDR $\leq \alpha$), then we assign the pattern with largest value of the test statistic. Thus, we are essentially adopting the ORIOGEN methodology developed in Peddada et al. (2003) to the present context.

5. NUMERICAL RESULTS

We evaluate the performance of our proposed methodology, which we refer to as ANCOM-II, using two distinct simulation studies. The first is inspired by a real data set collected by Yatsunenko et al. (2012). This setup also allows for all three kinds of zeros described in the paper. The second is based on a negative binomial distribution, which is commonly used to model OTU count data of microbiome studies. The results of the proposed method are obtained by filtering outlier zeros at a threshold of $c = 0.15$. We compare the proposed with methodology with three other methods, namely, DESeq2 (Love, Huber and Anders 2014), t-test based on sample proportions (Prop-T) and t-test based on data transformed via (2.3) after adding a pseudo-count of 1 to each entry (Pseudo-C). Note that a comparison between ANCOM-II and the Pseudo-C method provides numerical results on how our assessment of zeros impacts the analysis. We also provide a user friendly R code in the supplementary materials to implement the proposed methodology described in this section.

5.1. Simulation Study Based on Real Data

This simulation study is based on the OTU count data (at the genus level) corresponding to the US group provided in Yatsunenko et al. (2012). We constructed two groups, namely, cases and controls ($J = 2$). Each group consisting of 175 subjects and 200 taxa. Among these 200 taxa, 100 are taken to be differentially abundant. As detailed below, our simulation study allows for all three forms of zeros discussed in the paper.

Step 1 Generate a simple random sample of 175 subjects from the US group in Yatsunenko et al. (2012) data. Process the data as described in Section 2 by taking the genus *Bifidobacterium* as the reference taxon for the transformation (2.3). This provides us with a 175 × 661 data matrix. Let $m = (m_1, ..., m_{200})$ denote the vector of 200 column means which are highest in magnitude obtained after normalization of (2.3).

Step 2 (Outlier zeros) Using the vector m simulate 175 case and control samples using a bimodal distribution as follows. For $i = 1, .., 175$

$$\begin{aligned} y_{i1k} &\sim^{iid} \pi\mathcal{N}(m_k - 3, 1) + (1-\pi)\mathcal{N}(m_k + 3, 1), \\ &\quad k = 1, ..., 100 \\ y_{i2k} &\sim^{iid} \pi\mathcal{N}(m_k - 3, 1) + (1-\pi)\mathcal{N}(m_k + 3, 1), \\ &\quad k = 1, ..., 50 \\ y_{i2k} &\sim^{iid} \pi\mathcal{N}(m_k - 3, 1) + (1-\pi)\mathcal{N}(m_k + 3 + \delta, 1), \\ &\quad k = 51, ..., 100. \end{aligned}$$

For each simulated repetition π is chosen uniformly between (0.85,0.95).

Step 3 (Sampling zeros) Using the vector m simulate 175 case and control samples with a unimodal distribution. For $i = 1, .., 175$

$$\begin{aligned} y_{i1k} &\sim^{iid} \mathcal{N}(m_k, 1), \ k = 101, ..., 175 \\ y_{i2k} &\sim^{iid} \mathcal{N}(m_k, 1), \ k = 101, ..., 125 \\ y_{i2k} &\sim^{iid} \mathcal{N}(m_k + \delta, 1), \ k = 126, ..., 175 \end{aligned}$$

Step 4 (Structural zeros) Create 175 case and control samples for taxa that are structurally zero in the control group. For $i = 1, .., 175$, $k = 176, ..., 200$, set $y_{i1k} = 0$ and $y_{i2k} = \mathcal{N}(m_k, 1)$ with probability 0.01.

Step 5 Back transform the above continuous scale data to the count scale by inverting the transformation (2.3) and rounding the observations. Specifically, using the transformation

$$z_{ijk} = e^{y_{ijk}}\Big[z_{ijb}/\big(\prod_i z_{ijb}\big)^{1/n}\Big]$$

here z_{ijb} represents the counts of "*Bifidobacterium*" taxa in the subset of the global gut data described in **Step 1**. In the above steps, all values between (0,1) are rounded to zero counts. Thus, although we are generating continuous random variables, with a positive probability we generate zeros. Recall that in **Step 2** samples are generated from a mixture of two independent normal distributions. The observations corresponding to zero counts are induced by the first component of the mixture distribution. Since the two components are independently generated, the zero observations are not dependent on the taxa itself (assuming that the true distribution of the taxa is given by the second component). Thus, these zeros, by design, represent observations that are missing at random. On the other hand, the zeros obtained in **Step 3** are from a single distribution, and are zero because z_{ijk} with values between 0 and 1 are set to 0.

Step 6 Apply the three methods on the above simulated count data. Repeat Steps 1 through 6 and estimate the false discovery rate (FDR) and power of each method.

The left and right panels of **Figure 2** provides the estimated FDR and power of the four methods, respectively. Here the shift parameter of Steps 2 and 3 is set to $\delta = 0.5$. In this setting, on average (red dot), our proposed method, DESeq2 and Pseudo-C appear to control the FDR at the nominal level of 0.05. However, in terms of power our method appears to outperform the rest. In **Figure 3**, we further examine the effect of a varying shift parameter δ. We compare the powers of the four methods for 100 distinct values of $\delta \in (0, 0.5)$. Once again we note that the proposed method ANCOM-II, tends to have larger power than the others. Specifically, a comparison between ANCOM-II and the Pseudo-C method emphasizes the importance of identifying the various sources of zeros and dealing with them accordingly, rather than using a constant pseudo-count for all observed zeros.

5.2. Simulation based on Negative Binomial Distribution

In this section we investigate the performance of the four methods by generating data according to negative binomial (NB) distribution as follows. For $j = 1, 2$, $k = 1, ..., 200$, we generate,

$$z_{ijk} \sim NB(\mu_{jk}, s_{jk}), \quad i = 1, .., 100 \tag{5.1}$$

where μ_{jk}, s_{jk} are the mean and dispersion parameters of the negative binomial distribution respectively, in all cases we set $s_{jk} = \mu_{jk}^2$. The control samples are generated for $j = 1$ and $k = 1, .., 200$ by choosing μ_{jk} from a uniform distribution over (1,1500). The case samples are generated by shifting the mean of the first one hundred taxa. Thus,for $j = 2$, $k = 1, .., 100$ set $\mu_{jk} = \mu_{1k} + 5k$. The remaining $k = 101, ..., 200$ micorbes for group $j = 2$ are generated with the same mean parameters as the control samples. Furthermore we induce additional zeros in the data set by multiplying the previously generated counts with independent Bernoulli random variables $w_{ijk} = 0$ with probability $1 - \pi_{jk}$ where π_{jk} is chosen uniformly between (0.8,1). This simulation experiment is repeated 100 times and the FDR and power comparison results are reported in **Figure 4**. From these simulation results we note that only ANCOM-II and Pseudo-C have estimated FDR at or below the nominal level of 0.05. Furthermore, between the two methods, ANCOM-II enjoys higher power. DESeq2 and Prop-T have unacceptably high estimated FDR.

6. ANALYSIS OF GLOBAL HUMAN GUT MICROBIOME DATA

We illustrate ANCOM-II using global human gut microbiome data of Yatsunenko et al. (2012). The data consists of microbial taxa counts obtained from 317 subjects from US, 99 from Venezuela and 114 from Malawi. We used *Bifidobacterium* as the reference taxon because it was present in all samples.

Let S_i denote the set of genera with i countries having structural zeros. According to our method, by taking $c = 0.15$ we found that out of 661 genera, 262 belong to S_0, 86 belong to S_1, 95 belong to S_2 and 218 belong to S_3. Depending upon the set a genus belongs to, the method tests suitable hypotheses as outlined below (the corresponding R code is provided in the supplementary materials).

Hypotheses 1. For genera $j \in S_0$ we test the following hypothesis

$$\begin{aligned} H_{0j} &: \mu_{US,j} = \mu_{Venezuela,j} = \mu_{Malawi,j}, \quad \text{against} \\ H_{aj} &: \{\mu_{US,j} \le \mu_{Venezuela,j} \le \mu_{Malawi,j}\} \\ &\quad \cup \{\mu_{US,j} \le \mu_{Venezuela,j} \ge \mu_{Malawi,j}\} \\ &\quad \cup \{\mu_{US,j} \ge \mu_{Venezuela,j} \le \mu_{Malawi,j}\} \\ &\quad \cup \{\mu_{US,j} \ge \mu_{Venezuela,j} \ge \mu_{Malawi,j}\} \end{aligned}$$

Hypotheses 2a. For genera $j \in S_1$, when a taxon is structurally zero in Malawi data we test the following hypothesis

$$\begin{aligned} H_{0j} &: \mu_{US,j} = \mu_{Venezuela,j} \quad \text{against} \\ H_{aj} &: \{\mu_{US,j} \le \mu_{Venezuela,j}\} \\ &\quad \cup \{\mu_{US,j} \ge \mu_{Venezuela,j}\} \end{aligned}$$

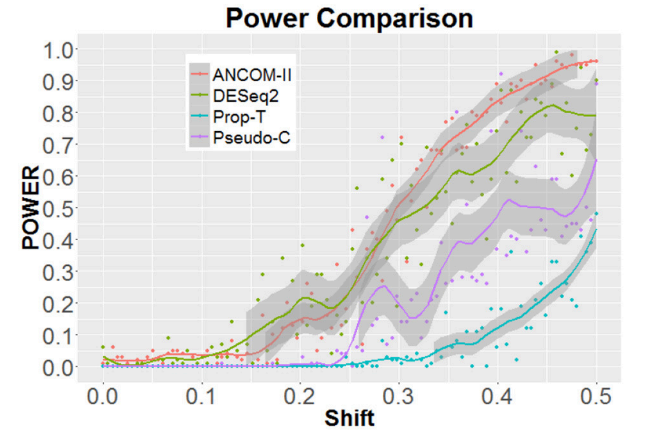


FIGURE 3 | Power comparisons among ANCOM II, DESeq2, Prop-T, and Pseudo-C, for different values of $\delta \in (0, 0.5)$.

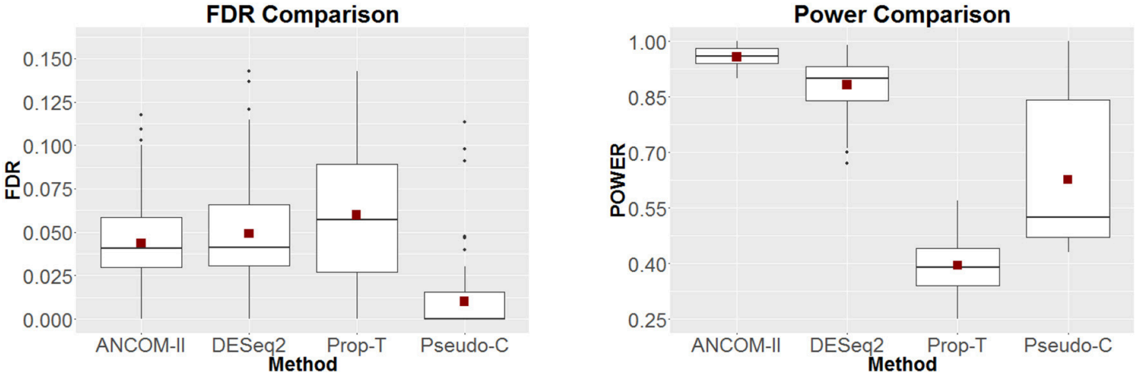


FIGURE 2 | FDR **(Left)** and Power **(Right)** comparisons among ANCOM II, DESeq2, Prop-T, and Pseudo-C. Power comparisons are for $\delta = 0.5$.

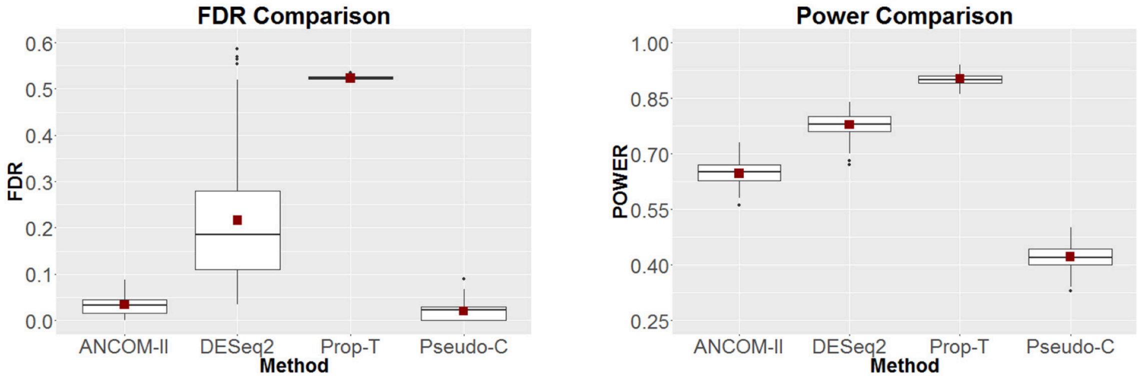


FIGURE 4 | FDR **(Left)** and Power **(Right)** comparisons among ANCOM II, DESeq2, Prop-T, and Pseudo-C for simulation based on negative binomial distribution.

Hypotheses 2b. For genera $j \in S_1$, when a taxon is structurally zero in Venezuela data we test the following hypothesis

$$\begin{aligned} H_{0j} &: \mu_{US,j} = \mu_{Malawi,j} \quad \text{against} \\ H_{aj} &: \{\mu_{US,j} \leq \mu_{Malawi,j}\} \\ &\quad \cup \{\mu_{US,j} \geq \mu_{Malawi,j}\} \end{aligned}$$

Hypotheses 2c. For genera $j \in S_1$, when a taxon is structurally zero in US data we test the following hypothesis

$$\begin{aligned} H_{0j} &: \mu_{Venezuela,j} = \mu_{Malawi,j} \quad \text{against} \\ H_{aj} &: \{\mu_{Venezuela,j} \leq \mu_{Malawi,j}\} \\ &\quad \cup \{\mu_{Venezuela,j} \geq \mu_{Malawi,j}\} \end{aligned}$$

Hypotheses 3. For genera $j \in S_2$, which is structurally zero in Malawi and Venezuela data, we declare it to be differentially abundant (relative to a reference taxon) in the US compared to the other two countries. A similar conclusion is arrived for the other two possibilities.

Hypotheses 4. All genera belonging to thisset are discarded because they are considered to be absent in all three data sets.

Using the above approach ANCOM-II, relative to *Bifidobacterium* identified a total of 83 differentially abundant genera. Furthermore, ANCOM-II identified patterns of relative abundance of genera over the three countries. For genera in set S_0 that are significant we discovered 34 genera belong to the phylum Firmicutes, followed by Proteobacteria (25), Actinobacteria (6), Tenericutes (5), Bacteroidetes (5) and others. Only 1 genera in set S_1 (absent in Malawi) was found significant and belonged to the phylum Proteobacteria. Numbers within parenthesis represent the number genera within each phylum that were significant. We note that, the second highest number of differentially abundant genera belonged to phyla Proteobacteria. This is surprising given that this is typically one of the smaller phyla in the gut microbiome. This phylum consists of a large number of opportunistic pathogenic bacteria and an increased abundance of Proteobacteria is known to be associated with the disease necrotizing enterocolitis (NEC) Wang et al. (2009); Mai et al. (2011) and Inflammatory Bowel Disease (IBD), [Balfour Sartor and Mazmanian (2012)]. The genera in this phylum were observed to be uniformly lower in the US group as compared to the other two. A total of 29 taxa were present in US but structurally zero in Venezuela and Malawi, 53 were present in Venezuela but structurally zero in US and Malawi, lastly 13 were present in Malawi but structurally zero in Venezuela and US. In addition to ANCOM-II, we also applied DESeq2, Prop-T and Pseudo - C methods to these data. The results are summarized in the Venn diagram provided in **Figure 5**.

For comparison purposes, we re-analyzed the data using ANCOM-II but using the geometric mean (GM) of all non-zero taxa within subject as the reference, instead of *Bifidobacterium*. All taxa identified using *Bifidobacterium* as the reference taxon were a subset of taxa identified by the geometric mean as the reference taxon. The results are summarized in the Venn diagram in **Figure 5**.

7. DISCUSSION

One of the challenges when dealing with compositional microbiome data is the presence of a large frequency of zero counts. At the moment there is no generally applicable methodology for comparing relative abundances of taxa among two or more populations/groups in presence of excess zero counts. In this article we took the first step toward identifying different types of zero counts and provided a strategy to deal with them. We take a principled approach to these data by classifying these zero counts into three different types. Inspired by gene expression studies, we proposed a simple method to "normalize" the data to eliminate specimen level effects. To deal with specimen specific background value, one may use a taxon that is present in all specimens, such as *Bifidobacterium* in the example considered in this paper, or one can use the geometric mean of taxa within the specimen. From our empirical studies, the choice of the background does not seem to affect the FDR, but could impact the power. Using this framework, a variety of statistical tests can be carried over from the literature depending upon the scientific question and hypotheses of interest. In this paper we describe four different types of statistical tests that are of common interest. Methodology

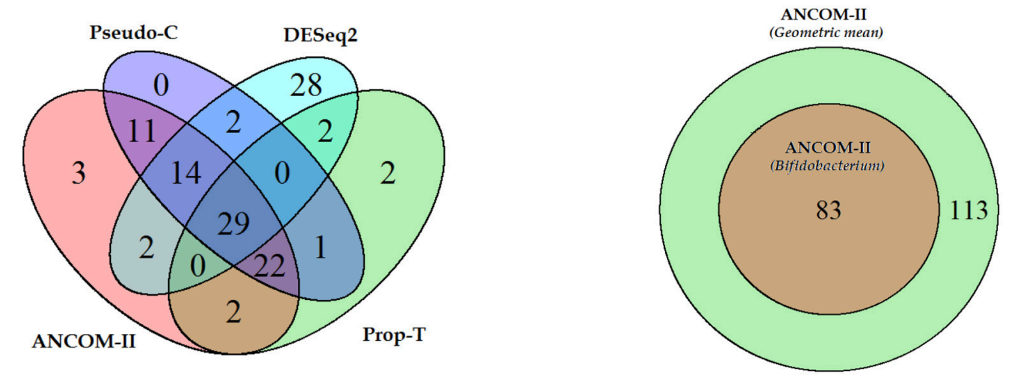


FIGURE 5 | (Left) Venn diagram illustrating overlapping features detected by different procedures. **(Right)** overlapping features detected by assuming *Bifidobacterium* as normalizer or the geometric mean of all taxa as the normalizer.

developed in this paper, called ANCOM - II, is a general procedure that is not only applicable to cross-sectional as well as longitudinal designs, but in each case it can be used for detecting trends and patterns in a taxon over two or more groups. Our simulation study suggests that the methodology controls the overall false discovery rate while maintaining high power. In addition, since the methodology is based on residual bootstrap, it does not make any major distributional assumptions. For testing non-directional alternative hypotheses (hypothesis H_1), ANCOM-II can be implemented using the R-code accompanying this paper. If no covariates are present and if there are no repeated measurements, then using residuals calculated in Equation (2.2) ANCOM-II can be implemented for testing directional alternatives H_2, H_3 by applying ORIOGEN. However, if covariates are present and if there are repeated measurements then ANCOM-II can be implemented for testing directional alternatives H_2, H_3 by applying CLME. At the moment we do not have a unified user friendly code that would be suitable for all scenarios described above. A general purpose software is being developed and we hope to release it in the near future.

AUTHOR CONTRIBUTIONS

AK: Conceived the ideas, developed methodology, performed all numerical work and edited the manuscript. SM: Conceived the ideas and edited the manuscript. OD: Conceived the ideas and edited the manuscript. SP: Conceived the ideas, developed methodology and edited the manuscript.

FUNDING

SP and AK were supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04). OD was partially supported by the Israeli Science Foundation Grant No. 1256/13 and Grant 456/17.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02114/full#supplementary-material

REFERENCES

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

Aitchison, J. (1985). A general class of distributions on the simplex. *J. R. Statist. Soc. B* 47, 136–146.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Chen, E. Z., and Li, H. (2016). A two-part mixed-effect model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617. doi: 10.1093/bioinformatics/btw308

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The Impact of the Gut Microbiota on Human Health: an Integrative View. *Cell* 148, 1258–1270. doi: 10.1016/j.cell.2012.01.035

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D. J., and Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res.* 54, 2325–2340. doi: 10.1194/jlr.R036012

Farnan, L., Ivanova, A., and Peddada, S. D. (2014). Constrained inference in biological sciences: linear mixed effects models under constraints. *PLoS ONE.* 9:e84778. doi: 10.1371/journal.pone.0084778

Grandhi, A., Guo, W., and Peddada, S. D. (2016). A multiple testing procedure for multi-dimensional pairwise comparisons with application to gene expression studies. *BMC Bioinformatics* 17:104. doi: 10.1186/s12859-016-0937-5

Guo, W., Sarkar, S. K., and Peddada, S. D. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* 66, 485–492. doi: 10.1111/j.1541-0420.2009.01292.x

Jelsema, C., and Peddada, S. D. (2016). *CLME: An R Package for Linear Mixed Effects Models under Inequality Constraints.* Journal of Statistical Software.

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

Mai, V., Young, C. M., Ukhanova, M., Wang, X., Sun, Y., Casella, G., et al. (2011). Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PLoS ONE* 6:e20647. doi: 10.1371/journal.pone.0020647

Mandal, S., Van Treuren, W., White, R. A., Eggesb, 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

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Peddada, S. D., Hwang, J. T. G. (2002). Classification of pixels in a noisy greyscale image of polar ice. *IEEE Trans. Geosci. Remote Sensing* 40, 1879–1884. doi: 10.1109/TGRS.2002.802517

Peddada, S. D., Lobenhofer, L., Li, L., Afshari, C., Weinberg, C., and Umbach, D. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* 19, 834–841. doi: 10.1093/bioinformatics/btg093

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581–592.

Sartor, R. B., and Mazmanian, S. K. (2012). Intestinal Microbes in Infl ammatory Bowel Diseases. *Am. J. Gastroenterol. Suppl.* 1, 15–21. doi: 10.1038/ajgsup.2012.4

Wang, Y., Hoenig, J. D., Malin, K. J., Qamar, S., Petrof, E. O., Sun, J., et al. (2009). 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J.* 3, 944–954. doi: 10.1038/ismej.2009.37

Xia, F., Chen, J., Fung, W., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* 69, 1053–1063. doi: 10.1111/biom.12079

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

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 Microbiology

ORIGINAL RESEARCH
published: 07 March 2018
doi: 10.3389/fmicb.2018.00297

Linking Associations of Rare Low-Abundance Species to Their Environments by Association Networks

Tatiana V. Karpinets[1,2]*, *Vancheswaran Gopalakrishnan*[3,4], *Jennifer Wargo*[1,3], *Andrew P. Futreal*[1], *Christopher W. Schadt*[2,5] *and Jianhua Zhang*[1]

[1] Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3] Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [4] Department of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Dallas, TX, United States, [5] Department of Microbiology, University of Tennessee, Knoxville, Knoxville, TN, United States

OPEN ACCESS

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Rohita Sinha,
University of Nebraska-Lincoln,
United States
Stephen Woloszynek,
Drexel University, United States

***Correspondence:**
Tatiana V. Karpinets
tvkarpinets@mdanderson.org

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

Received: 20 October 2017
Accepted: 08 February 2018
Published: 07 March 2018

Citation:
Karpinets TV, Gopalakrishnan V,
Wargo J, Futreal AP, Schadt CW
and Zhang J (2018) Linking
Associations of Rare Low-Abundance
Species to Their Environments by
Association Networks.
Front. Microbiol. 9:297.
doi: 10.3389/fmicb.2018.00297

Studies of microbial communities by targeted sequencing of rRNA genes lead to recovering numerous rare low-abundance taxa with unknown biological roles. We propose to study associations of such rare organisms with their environments by a computational framework based on transformation of the data into qualitative variables. Namely, we analyze the sparse table of putative species or OTUs (operational taxonomic units) and samples generated in such studies, also known as an OTU table, by collecting statistics on co-occurrences of the species and on shared species richness across samples. Based on the statistics we built two association networks, of the rare putative species and of the samples respectively, using a known computational technique, Association networks (Anets) developed for analysis of qualitative data. Clusters of samples and clusters of OTUs are then integrated and combined with metadata of the study to produce a map of associated putative species in their environments. We tested and validated the framework on two types of microbiomes, of human body sites and that of the *Populus* tree root systems. We show that in both studies the associations of OTUs can separate samples according to environmental or physiological characteristics of the studied systems.

Keywords: metagenome, microbiome, unsupervised analysis, alpha and beta diversity, sparse data, Anets, qualitative data

INTRODUCTION

The rare low-abundance microbial species, which have been referred to as the "rare biosphere" (Sogin et al., 2006), have attracted increasing attention in the recent literature because of their unknown ecology and potential evolutionary and ecological importance (Youssef et al., 2010; Pedros-Alio, 2012; Coveley et al., 2015; Lynch and Neufeld, 2015; Sharon et al., 2015; Jousset et al., 2017). Although sequencing errors and undersampling of OTUs may contribute to extent of the "rare biosphere," the advent of new bioinformatics tools (Schloss and Westcott, 2011; Preheim et al., 2013; Edgar and Flyvbjerg, 2015; Sharon et al., 2015; Callahan et al., 2016) as well as experimental and technological approaches (Jousset et al., 2017) are increasingly compelling of the presence and complexity of these rare taxa. Biological explanations (Pedros-Alio, 2012; Coveley et al., 2015; Lynch and Neufeld, 2015; Jousset et al., 2017) and other factors, such as poor taxonomic resolution of short reads, especially for closely related species or those poorly represented in the genomic database, incomplete or inadequate sampling, dispersal limitation, spatial and temporal partitioning of the environment, and the nestedness of ecological mutualistic networks, may contribute to such results (Bascompte et al., 2003; Youssef et al., 2010; Rosindell et al., 2011; Unterseher et al., 2011; James et al., 2012; Mi et al., 2012; Pedros-Alio, 2012; Suweis et al., 2013).

The numerous rare OTUs are a typical output of 16S rRNA amplicon sequencing studies, especially those with many and diverse samples. The resultant sparse datasets present a challenge for common statistical tools. The data matrix produced by such studies are usually comprised of species-like groups (rows) and their abundances calculated as the number of sequencing reads representing each species across multiple samples (columns). The species-like groups are typically inferred by a conventional aggregation of sequences into OTUs based on a sequence identity threshold or, in more recent work, by amplicon sequence variants (ASVs) (Callahan et al., 2016; Callahan, 2017). In both cases, most species-like groups could be representative of species-specialists; they are not only low in abundance in a given sample, but are also rare across samples and environments. Known computational tools for analyzing the sparse data often address the sparsity problem by filtering out very rare species or by collapsing species to a higher-level hierarchy. Although the aggregation reduces sparsity (dominance of zeros in the dataset) of the data, the OTUs-level insights into the structure of microbiome will be lost. By excluding the rare OTUs, such as those found in less than 30% of samples, we also may lose information. It is not clear how extensive this loss might be.

In addition to sparsity, the 16S rRNA gene sequencing data have other challenges including their compositionality and dimensionality (essentially greater number of OTUs than the number of samples). The data compositionality means that we don't know the real OTU abundances and have to deal with proportions of species relative to their sum in each sample. Several methods have been proposed to address the challenges (McMurdie and Holmes, 2014; Tsilimigras and Fodor, 2016). The most recent methods proposed to infer species–species relationships from the 16S rRNA amplicon datasets include Compositionality Corrected by REnormalization and PErmutation (CCREPE) (Faust et al., 2012), metagenomeSeq (Paulson et al., 2013), Sparse Correlations for Compositional data (SparCC) (Friedman and Alm, 2012), a mixture model framework (McMurdie and Holmes, 2014), SParse InversE Covariance Estimation for Ecological Association Inference (SpiecEasi) (Kurtz et al., 2015), and gCoda (Fang et al., 2017). Each of the tools addresses dimensionality and compositionality challenges of the datasets using different computational approaches. The cumulative sum scaling normalization and the zero-inflated Gaussian distribution mixture model are used in metagenomeSeq to account for biases resulting from under-sampling when selecting the differential abundant OTUs. The log-ratio transformation and the variance are used in SparCC to overcome compositionality of the data. The data dimensionality and compositionality are even more efficiently addressed by SpiecEasi and gCoda using the data transformation borrowed from the compositional data analysis and then inferring the interaction graph from the transformed data by neighborhood selection or by sparse inverse covariance selection.

All abovementioned tools, however, analyze the OTU table after filtering out most rare OTUs (Supplementary Figures S1A–D). In case of SparCC, the filtering is the most stringent because the algorithm employs log-transformations of the read counts. The basic assumption of the approach is that all OTUs are present in the dataset; therefore small values must be assigned to undetected OTUs to include them in the analysis. The percentage of rare OTUs may be even greater in studies with large number of samples or when sampling takes place in more diverse environments, such as the Human Microbiome Project (HMP) dataset and the *Populus* Root Microbiome (PRM) dataset (Supplementary Figures S1E,F). In the study we have made an attempt to explore the biological role of the rare low-abundance OTUs in these two environments using existing data from Human body sites (2012) and from *Populus* roots (Shakya et al., 2013). To reduce the burden of filtering for the rare OTUs and overcome the problem of compositionality we treat the OTUs as qualitative variables and apply an analytical tool specific for analysis of such datasets.

RESULTS

Approach

Our initial analysis of the Human and *Populus* microbiome datasets reveals that both datasets are in agreement with the well-known occupancy–abundance relationship (Gaston, 1996), which positively links the species abundances and the number of sites/samples they occupy. We find that in both datasets, OTUs that are more common across samples are also more abundant, and rare OTUs across samples are usually less abundant (**Figures 1A,B**). Notably, the number of common abundant OTUs is extremely small in the datasets. Considering this observation we decided to treat the rare OTUs as qualitative data by replacing the putative species abundances with the presence/absence call (0/1 values). Although in this approach we lose information on abundances, at the same time, the resulting dataset will not be compositional. In addition, we get the chance to transform the data to collect additional statistics on co-occurrences of species with each other and to quantify interdependencies of the species. The quantification is based on an assumption that rare OTUs (putative species) are associated because they are dependent upon one another in each studied environment. They may be dependent metabolically, when metabolites produced by one species are consumed by another species. They also may have similar optimal growth conditions or offer complementary functions to support microbial community as a whole (Jousset et al., 2017). All these factors may lead to co-occurrences of the rare OTUs in the samples. We quantify the co-dependence of OTUs by calculating a co-occurrence profile of each OTU with all other OTUs in the data and by interrogating

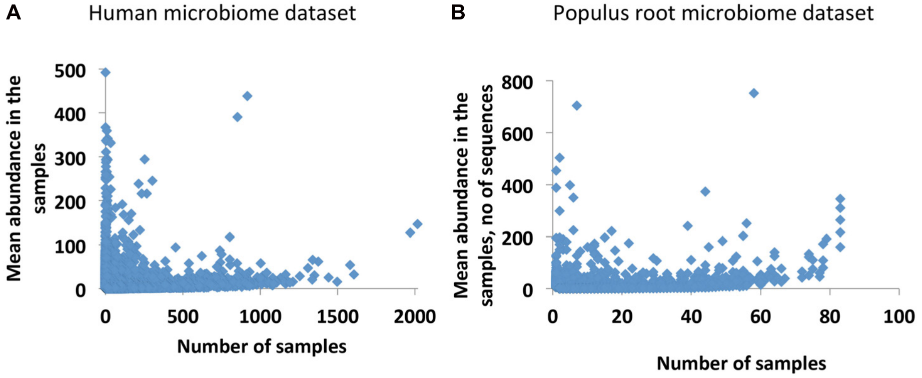


FIGURE 1 | Occupancy–abundance relationship. **(A)** Human Microbiome Project (HMP) dataset (43140 OTUs × 2910 Samples). **(B)** *Populus* Root Microbiome (PRM) dataset (24434 OTUs × 83 Samples).

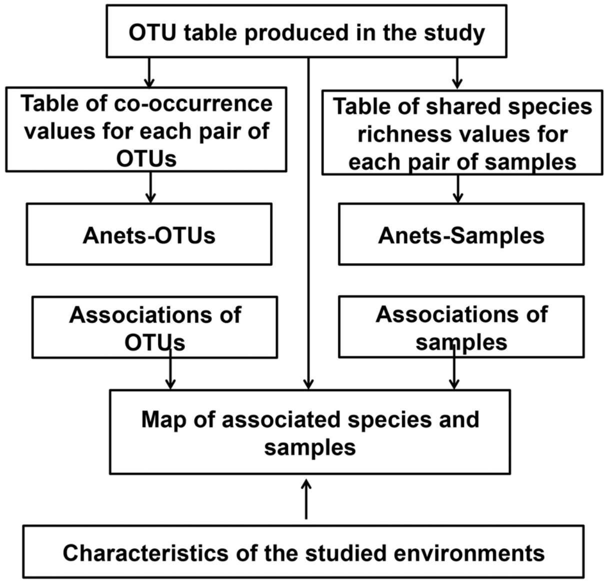

OTU table produced in the study
Table of co-occurrence values for each pair of OTUs
Table of shared species richness values for each pair of samples
Anets-OTUs
Anets-Samples
Associations of OTUs
Associations of samples
Map of associated species and samples
Characteristics of the studied environments

FIGURE 2 | Computational framework used in the study to explore associations of rare species.

similarities of the emerged profiles for each pair of OTUs. We performed the calculations by applying a previously developed statistical tool, Association Network (Anets) (Karpinets et al., 2012), used for discovering of associations in qualitative datasets[1] and refer to the resultant network as Anets-OTUs.

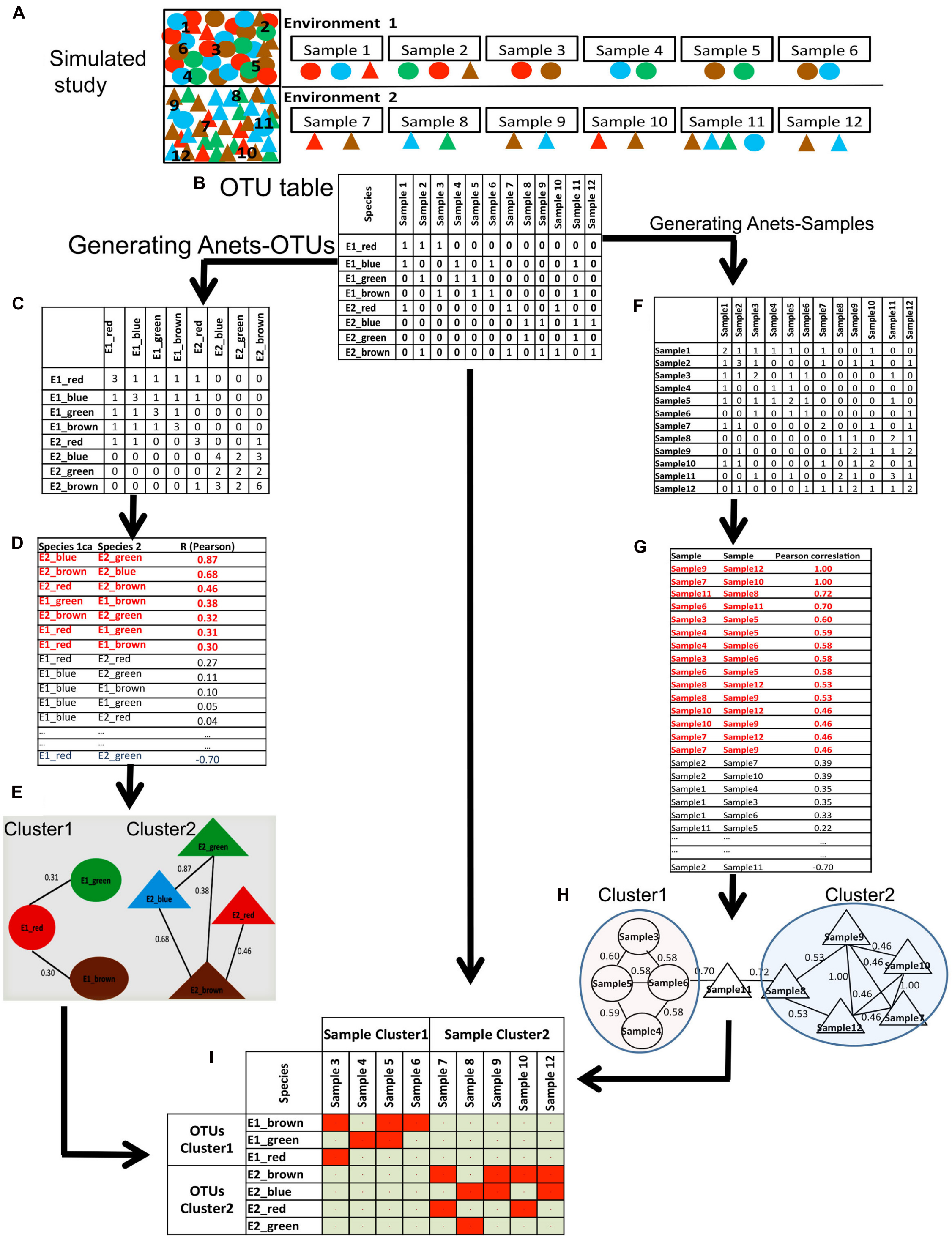
In addition to that Network, we also build the network of samples, Anets-Samples, using the same algorithm. By combining both networks we produce a map where associated OTUs and associated samples are clustered according to their presence/absence. This map can be further compared with characteristics of the studied environments. An overview of this computational framework is shown in **Figure 2** and details of the implementation are provided in Supplementary Data Sheet S1.

We also used a simulated dataset (**Figure 3A**) to illustrate and explain computations underlying the proposed framework. In this study, we have two synthetic microbial communities with four associated species (circles) in the first community and four associated species (triangles) in the second community. Species in each community are co-dependent, and therefore more often co-occur in their parent environment. We made 12 random samples of species from the communities and organized the sampling results as an OTU table (**Figure 3B**) with species/OTUs in rows and samples in columns. All species identified in the samples are rare; they are found only in 2–5 out of 12 samples. Thus, we replaced the species abundances with the presence/absence (1/0) values.

Association Network of Species

To generate the Anets-OTUs we first transform the OTU table to produce a new table where rows and columns consist of OTUs and each cell shows the number of samples where two OTUs co-occur in the data (**Figure 3C**). The transformed table, therefore, gives us the co-occurrence profiles for each OTU with the rest. We further use these profiles to infer pair-wise associations of the OTUs (**Figure 3D**). Although the input of the approach is OTU table with 1/0 values instead of counts, the statistics collected in the transformed table produces continuous variables. The Anets program provides three options to quantify the pair-wise similarities of the profiles. The options include Spearman correlation (default), Pearson correlation, and cosine (Jaccard index). While alternative similarity metrics may be appropriate for particular datasets, in these studies we found that the Pearson correlation coefficient was most robust for identifying association networks. We calculate the Pearson correlation to measure similarity of the profiles for each pair of OTUs and consider the OTUs associated if the correlation coefficient R > = 0.30. The selected pairs of OTUs predict the network (Anets-OTUs) of seven species with seven associations separated into two clusters/communities (**Figure 3E**). The species inferred by the Anets-OTUs in each cluster correspond to two communities provided in the mock study (**Figure 3A**). The algorithm did not recover only one species from the Environment 1 of the study.

[1] https://sourceforge.net/projects/Anets/



B OTU table

Species	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12
E1_red	1	1	1	0	0	0	0	0	0	0	0	0
E1_blue	1	0	0	1	0	1	0	0	0	0	1	0
E1_green	0	1	0	1	1	0	0	0	0	0	0	0
E1_brown	0	0	1	0	1	1	0	0	0	0	1	0
E2_red	1	0	0	0	0	0	1	0	0	1	0	0
E2_blue	0	0	0	0	0	0	0	1	1	0	1	1
E2_green	0	0	0	0	0	0	0	1	0	0	1	0
E2_brown	0	1	0	0	0	0	1	0	1	1	0	1

C

	E1_red	E1_blue	E1_green	E1_brown	E2_red	E2_blue	E2_green	E2_brown
E1_red	3	1	1	1	1	0	0	0
E1_blue	1	3	1	1	1	0	0	0
E1_green	1	1	3	1	0	0	0	0
E1_brown	1	1	1	3	0	0	0	0
E2_red	1	1	0	0	3	0	0	1
E2_blue	0	0	0	0	0	4	2	3
E2_green	0	0	0	0	0	2	2	2
E2_brown	0	0	0	0	1	3	2	6

D

Species 1ca	Species 2	R (Pearson)
E2_blue	E2_green	0.87
E2_brown	E2_blue	0.68
E2_red	E2_brown	0.46
E1_green	E1_brown	0.38
E2_brown	E2_green	0.32
E1_red	E1_green	0.31
E1_red	E1_brown	0.30
E1_red	E2_red	0.27
E1_blue	E2_green	0.11
E1_blue	E1_brown	0.10
E1_blue	E1_green	0.05
E1_blue	E2_red	0.04
…	…	…
…	…	…
E1_red	E2_green	-0.70

F

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	Sample11	Sample12
Sample1	2	1	1	1	1	0	1	0	0	1	0	0
Sample2	1	3	1	0	0	0	1	0	1	1	0	1
Sample3	1	1	2	0	1	1	0	0	0	0	1	0
Sample4	1	0	0	1	1	0	0	0	0	0	0	0
Sample5	1	0	1	1	2	1	0	0	0	0	1	0
Sample6	0	0	1	0	1	1	0	0	0	0	0	1
Sample7	1	1	0	0	0	0	2	0	0	1	0	1
Sample8	0	0	0	0	0	0	0	1	1	0	2	1
Sample9	0	1	0	0	0	0	0	1	2	1	1	2
Sample10	1	1	0	0	0	0	1	0	1	2	0	1
Sample11	0	0	1	0	1	0	0	2	1	0	3	1
Sample12	0	1	0	0	0	1	1	1	2	1	1	2

G

Sample	Sample	Pearson correslation
Sample9	Sample12	1.00
Sample7	Sample10	1.00
Sample11	Sample8	0.72
Sample6	Sample11	0.70
Sample3	Sample5	0.60
Sample4	Sample5	0.59
Sample4	Sample6	0.58
Sample3	Sample6	0.58
Sample6	Sample5	0.58
Sample8	Sample12	0.53
Sample8	Sample9	0.53
Sample10	Sample12	0.46
Sample10	Sample9	0.46
Sample7	Sample12	0.46
Sample7	Sample9	0.46
Sample2	Sample7	0.39
Sample2	Sample10	0.39
Sample1	Sample4	0.35
Sample1	Sample3	0.35
Sample1	Sample6	0.33
Sample11	Sample5	0.22
…	…	…
…	…	…
Sample2	Sample11	-0.70

I

		Sample Cluster1				Sample Cluster2				
	Species	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Sample 12
OTUs Cluster1	E1_brown	🟥	·	🟥	🟥	·	·	·	·	·
	E1_green	·	🟥	🟥	·	·	·	·	·	·
	E1_red	🟥	·	·	·	·	·	·	·	·
OTUs Cluster2	E2_brown	·	·	·	·	🟥	·	🟥	🟥	🟥
	E2_blue	·	·	·	·	·	🟥	🟥	·	🟥
	E2_red	·	·	·	·	🟥	·	·	🟥	·
	E2_green	·	·	·	·	·	🟥	·	·	·

FIGURE 3 | Generating Anets-OTUs using the simulated study. **(A)** A simulated study of two synthetic microbial communities: four species shown by colored (red, green, blue, brown) circles (Community 1), and four different species shown by colored (red, green, blue, brown) triangles (Community 2). The same color of the species indicates their close taxonomic relationship. To introduce noise in sampling, two species from the second community were added to the first community, and one species from the first community was added to the second community. Six samples were taken to identify species in each community and to generate an OTU table with the species abundances. **(B)** OTU table of the simulated study. **(C)** The table of co-occurrences for each pair of OTUs. Values of the table show the number of samples where each pair of species co-occurs. **(D)** Pair-wise similarities of the co-occurrence profiles for each pair of species. Red colored associations were used to generate Anets-OTUs. **(E)** Anets-OTUs. **(F)** The table of the shared species richness for each pair of samples. Values of the table show how many OTUs are shared for each pair of samples. **(G)** Pair-wise similarities of the shared species richness profiles for each pair of samples. Red colored associations were used to generate Anets-Samples. **(H)** Anets-Samples. **(I)** A map of the associated species and samples.

While, the calculations described in this small illustrative dataset can be implemented in Excel, in case of real datasets, with many samples and OTUs, the calculations can be performed using the Anets program (Karpinets et al., 2012). The program also calculates the *p*-value for each association using the Monte-Carlo simulation. The associated species, therefore, can be selected using a *p*-values threshold. The Anets-OTUs produced for the mock study is small and doesn't require clustering. For the real dataset, different algorithms and software tools can be used to cluster the network as described in Supplementary Data Sheet S1.

Association Network of Samples

A similar algorithm was used to generate the associations of samples (**Figures 3F–H**). In this case we transform the OTU table to produce a new table where both rows and columns consist of samples and each cell represents the number of shared OTUs for each pair of samples. The ecological interpretation of the number is the shared species richness for a pair of samples. We consider two samples associated if they have a similar profile of the shared species richness values across all samples in the dataset. Such indirect similarity can establish an association between each pair of samples even if the majority of species in the samples are not common. Computationally, the algorithm generating the Anets-Samples (**Figures 3F–H**) is similar to the algorithm of the Anets-OTUs (**Figures 3C–E**). As before, the transposed table is used to compute profiles of shared species richness values for the samples (**Figure 3F**) followed by estimation of pair-wise correlations (**Figure 3G**) and clustering (**Figure 3H**). As we can see in the **Figure 3H**, the clustering recovers associations among 9 out of 12 samples in the illustrative study. The final step of the framework is an integration of the results obtained by Anets-OTUs and Anets-Samples by building a presence/absence map of the associated species and samples (**Figure 3I**).

Applying the Approach to Experimental Datasets

In order to test our methodology, we employed the described framework to analyze two well- established and published experimental datasets from a study of Human Microbiome Project Consortium, 2012 and from a study of the PRM (Shakya et al., 2013). In each of these datasets, 16S or 28S rRNA amplicon sequencing was used to profile the microbiome in different environments. By applying our methodology in an unsupervised manner to build a map of associated OTUs and samples, we were able to test how well the inter-sample associations reproduced their observed phenotype in the environment, with the added advantage of studying associations of rare OTUs underlying the grouping of samples.

Populus Root Microbiome

The dataset (Shakya et al., 2013) includes 2999 fungal OTUs and 24435 bacterial OTUs identified in 84 samples taken in May and in September from two geographical locations, Tennessee (TN) and North Carolina (NC) associated with the roots of Eastern Cottonwood (*Populus deltoides)* trees at along two different rivers. The study also collected a set of soil properties and host characteristics for each of the 23 sampling locations; we used these metadata to examine their relationships with the associations of samples discovered by the Anets-Samples.

Examination of the OTU table from the study reveals that common species (found in ~60% of samples) or generalists in *Populus* root are represented by only 61 OTUs, or 0.22% of total number of OTUs in the dataset. As expected, the majority of OTUs had low-abundance and was rare (**Figure 1B**). After applying the Anets-OTUs algorithm to the OTU table we found six large associations of OTUs (p-value < 0.05). A further enrichment analysis (see section "Materials and Methods") attributed each association to a location, TN and NC, and to a sampling season, May or September (**Figure 4A**). This analysis revealed that communities of low-abundance OTUs, were underlying groups of samples based on known environmental factors from the study. To further confirm the grouping we built a heat map of the associated OTUs (horizontal axis) across all samples (vertical axis) organized by the geographical location and season and sampling (**Figure 4B**). While, it can be appreciated that many rare Anets-OTUs are present across all samples, some of them often co-occur in samples from a particular location or a season. The largest microbial association includes OTUs found in *Populus* rhizosphere in any season and in any location. Some associations are more common for TN or NC, and some associations are more common in September or May. This pattern suggests a tight link between the identified associations of the rare OTUs and a particular environmental factor. We noticed, for example, that a fungal OTU representing the genus *Inocybe* was found only in the NC cluster. Indeed, species of the genera have been tied to their environments rather than their hosts more than other fungal species (Cripps, 1997). Our results are consistent with this experimental observation; they also indicate that the other fungal genera in the cluster, such as *Ceratobasidium*, have similar biological characteristics.

The analysis confirms that clustering at low taxonomic levels may be crucial in discriminating different environments. We find that although OTUs in each of the associations often belong to the same phyla, they are more distinct at lower taxonomic levels, such as order (Supplementary Tables S1A,B). For example, microbial communities of *Populus* roots in both locations, TN and NC, include phylums *Proteobacteria* with less number of OTUs in NC (Supplementary Table S1A). At the level of order, however, the *Proteobacteria* in NC had greater richness (10 orders) when compared with TN (seven orders), and included *Rhodocyclales, Syntrophobacterales, Rhodobacterales,* and *Burkholderiales* orders that were not observed in TN. Microbial communities in both locations, TN and NC, also included numerous species from phylum *Acidobacteria.* The microbial community in TN, however, was dominated by the order Solibacterales; this taxa, however, was not found in NC. This example clearly demonstrates that by analyzing the dataset at the level of OTUs and collapsing them after linking their associations to environments may be a better strategy for exploration of subtle difference among microbiomes in similar environments.

By applying the Anets-Samples algorithm to the OTU table we revealed two distinct clusters of samples in the PRM dataset

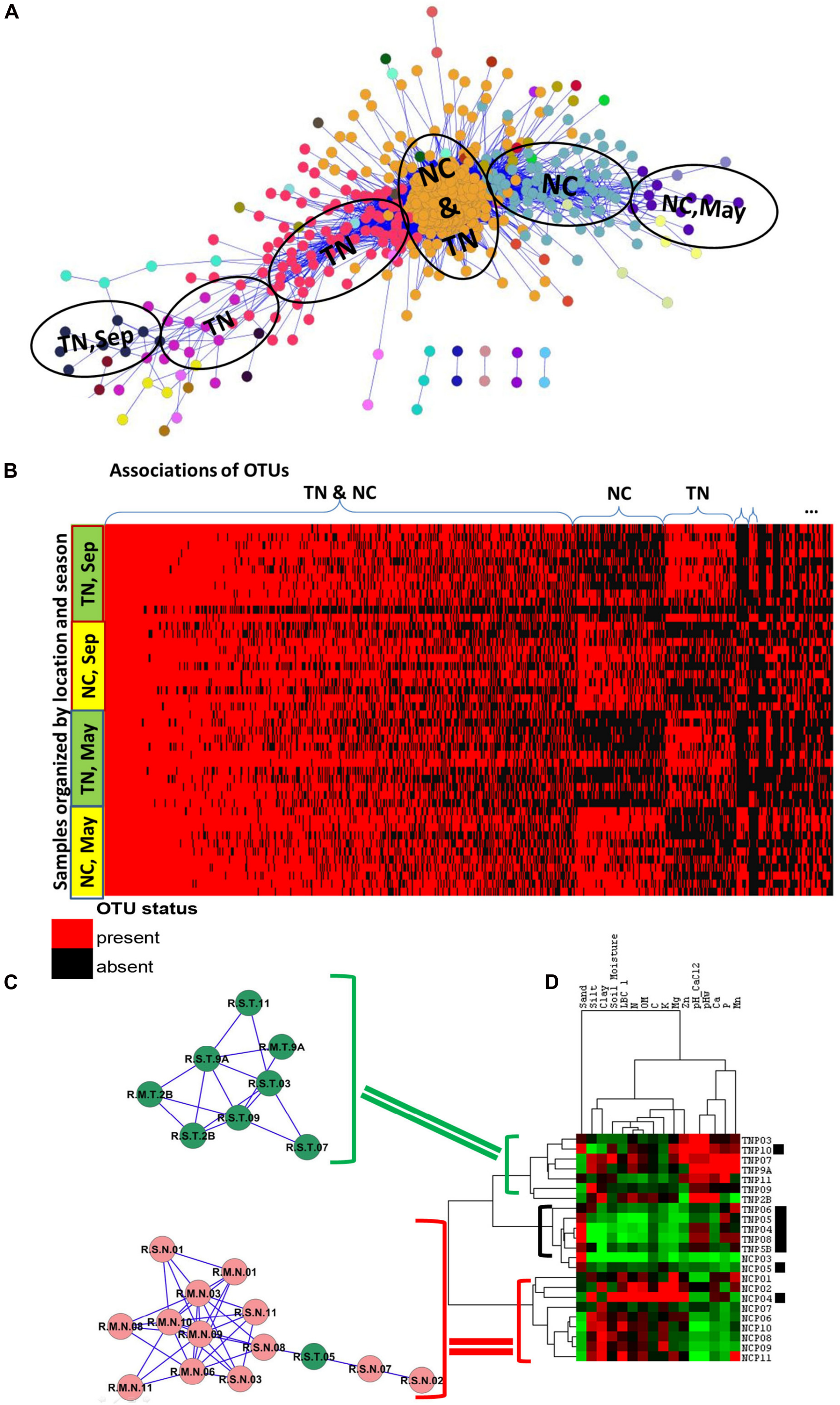


FIGURE 4 | Associations of rare species and samples in PRM study. **(A)** Communities of associated fungal and bacterial OTUs discovered by the Anets-OTUs algorithm in rhizoshpere of *Populus deltoides*. Nodes in the network indicate OTUs and edges indicate pair-wise association between them. The node color shows the community (cluster) assignment inferred by clustering**.** **(B)** Presence–absence map of the associated OTUs; the cell color is red if OTU is present in the sample and it is black if OTU is absent. OTUs are grouped according to the microbial communities inferred by Anets-OTUs and sorted by mean abundance; samples are grouped according to clusters inferred by Anets-Samples and sorted by the shared richness. **(C)** Two associations of *Populus* rhizoshpere samples with the shared species richness revealed by Anets-Samples; color indicates samples taken in NC (red) and in TN (green). **(D)** Hierarchical clustering of the soil properties; brackets indicate three cluster of soil samples with distinct soil properties: green bracket indicates the cluster of soil samples that correspond to the association of rhizosphere samples in TN, red bracket indicates the cluster of soil samples that correspond to the association of rhizosphere samples in NC, black bracket and black squares indicate samples that don't found as associated by Anets-Samples.

(**Figure 4C**). Within each cluster, all samples had similar profiles of the shared species richness across all samples ($p < 0.01$). Furthermore, there was a clear association with metadata of the study, with the first cluster representing a subset of samples from TN, and the second cluster representing a subset of samples from NC. Eight samples did not associate with either cluster. These results mirror the results of Shakya et al. (2013) that used variance partitioning of transformed datasets to show that watershed (TN vs. NC), season, and sampling site within a watershed, respectively, had the greatest effect on community structure followed by other factors. To determine other environmental factors contributing to the separation of samples in two clusters we examined the variance partitioning of the bacterial OTUs within each cluster with respect to host and soil properties, geographic locations, seasons, and diversity of corresponding fungal community. The analysis was performed the same way as in the original study (see section "Materials and Methods"). A large proportion of variance (67.8%) of the bacterial OTUs across all samples was unexplained in the original study, whereas only 9% of variance was explained by soil properties. In contrast, among the samples that were selected by the Anets-Samples as significantly associated, only 25% of variance remained unexplained, while the greatest proportion of the variance (30.1%) was attributed to the studied soil properties (Supplementary Figure S2). The expected proportion of the variance estimated by the permutation test, via a random selection of the same number of samples, would be only 19%.

To examine the effect of soil on the separation of samples in more detail we hierarchically clustered 16 soil properties measured in the study and found that two associations discovered by the Anets-Samples in *Populus deltoides* rhizosphere (**Figure 4C**), correspond to two distinct soil clusters inferred from the soil properties (**Figure 4D**). This relationship was not found in the original study and again suggesting the importance of rare microbial species for differentiating subtle environmental conditions in addition to the traditional methods that more heavily weight species abundance and dominant taxa. In case of PRM we observe that a set of TN samples found as associated by Anets share relatively greater Zn, Mn, and Ca contents in the soil and a greater soil pH. A set of associated NC samples share relatively low values of these soil characteristics. Those samples, either from TN or NC, that are not identified by Anets-Samples as significantly associated, have a variable content of the soil properties as well as relatively greater sand content and lower clay and organic matter contents than the associated samples. The results point to the soil properties as a crucial factor underlying similarity of microbial communities in *Populus deltoides* rhizosphere.

Microbiomes of Human Body Sites

The HMP dataset has been characterized in several publications (Faust et al., 2012; Project, 2012; Aagaard et al., 2013) and includes samples obtained from 18 different body sites of 180 healthy men and women. As noted before (**Figure 1A**), the majority of OTUs in the dataset is rare and has low-abundance. Considering the large size of the OTU table produced in the study we started the analysis with the construction of the Anets-Samples (**Figures 3F–H**) to find associations (clusters) of samples with similar profiles of the shared species richness and to discard samples-outliers. Most samples (74%) in the dataset were found to be associated (p-value < 0.01) with at least one other sample in the network. Visualization and clustering of the network using the Markov clustering algorithm (MCL) (Van Dongen, 2008) revealed seven large disconnected component and 206 clusters (Supplementary Figure S3). We next used an enrichment analysis (see section "Materials and Methods") to annotate the inferred clusters by sample metadata (sex of the human subject, body site, and sub-site) and to assign significantly enriched body sites and sub-sites to the clusters. **Figure 5A** shows components of the network comprised of oral and skin samples colored according to sub-sites. Samples that belonged to a particular subsite tended to cluster together according to the Figure and to the enrichment analysis. Thus, the Anets-Samples allowed us to predict origin of samples from different oral sub-sites, such as keratinized gingiva, buccal mucosa, hard palate, saliva, throat, and tongue. There were also several distinct associations of samples originated from multiple skin subsides. Interestingly, one association of samples (cluster 16 in **Figure 5A**) was comprised of male human subjects.

We further focused the analysis on 314 skin samples that represent three distinct, disconnected in the Anets-Samples, clusters labeled by black ovals in **Figure 5A**. To reveal communities of microbial OTUs discriminating these clusters we built the Anets-OTUs using, as input, an OTU table comprised of these 314 samples in columns and 43140 OTUs in rows. The generated Anets-OTUs included 412 associated OTUs (p-value < 0.001); and subsequent clustering of the network revealed four major microbial communities (**Figure 5B**). The enrichment analysis showed statistically significant links between the communities and the Anets-Samples clusters (Supplementary Table S2). The map generated from the initial dataset by extracting abundance values of the associating OTUs further confirmed the links (**Figure 5C**). Importantly, the three distinct clusters of samples, originated from skin of different human subjects, have significant differences in microbial communities at the OTU level, although most OTUs contributing to the difference belonged to the genus *Propionibacterium.* Indeed, microbial community 1 comprised of OTUs of the genus *Propionibacterium* (**Figure 5B**) was significantly enriched in Anets-Samples clusters 2 and 10 (**Figure 5A**), but not in Anets-Samples cluster 16 (**Figure 5A**). Microbial community 2 comprised of a distinct set of OTUs from the same genus (**Figure 5B**) was significantly enriched only in Anets-Samples cluster 2 (**Figure 5A**). The third microbial community comprised of OTUs of the genera *Propionibacterium* and *Actinomycetales* (**Figure 5B**) was enriched in Anets-Samples cluster 10 (**Figure 5A**), and the fourth microbial community (OTUs from the genera *Staphylococcus* and *Propionibacterium*) was enriched in Anets-Samples cluster 16 comprised of male human subjects. The p-value 0.01 (Fisher exact test) was used as the significance threshold in the enrichment analysis. Thus, the OTU level clustering was important to discriminate microbial communities of the clustered samples.

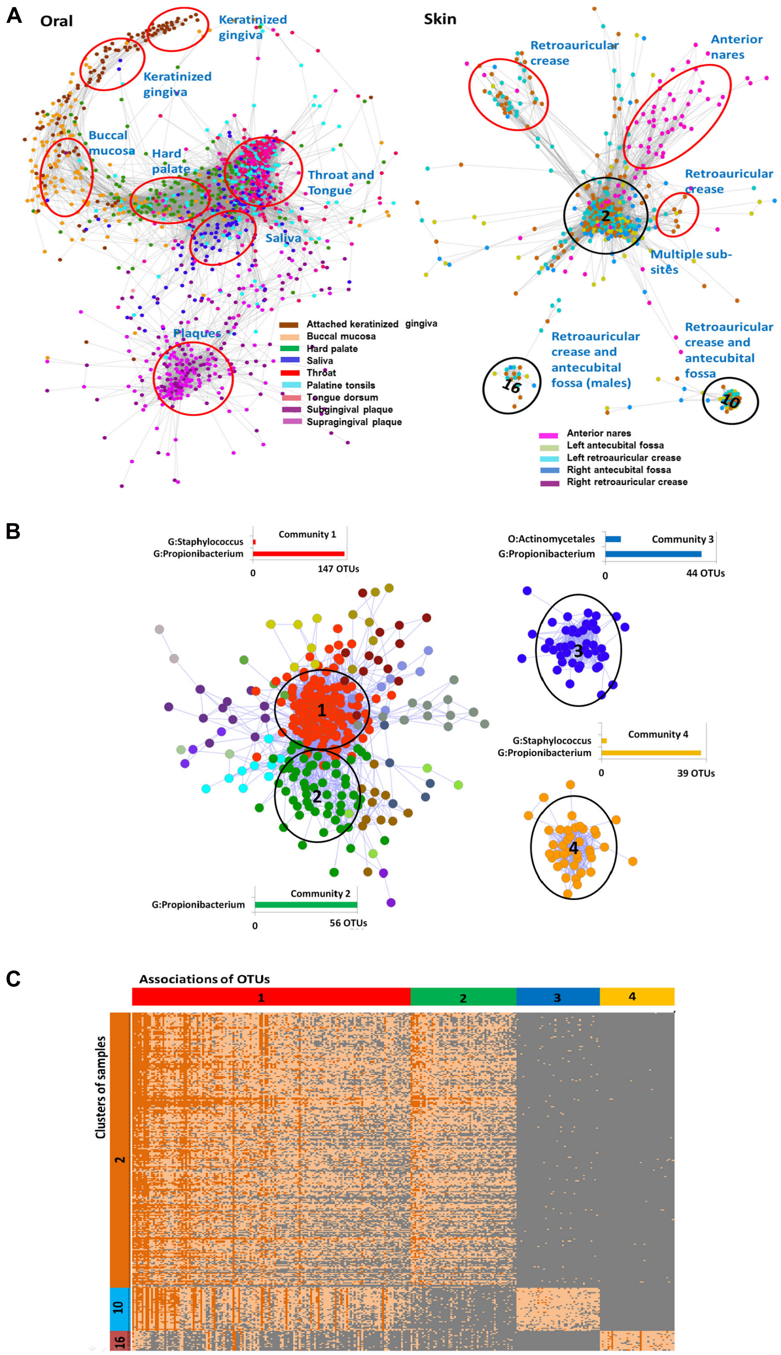


FIGURE 5 | Associations of rare species and samples in the HMP study. **(A)** Associations of oral and skin samples. Samples in the networks are represented by filled circles colored according to the sampling sub sites in the HMP study. Edges between circles indicate significant association between samples in terms of the shared species richness. Red and black ovals label associations predicted by clustering of the Anets-Samples. Name of each cluster was inferred by the enrichment analysis as described in Section "Materials and Methods." Black ovals indicate clusters (2, 10, and 16) that were further analyzed by the Anets-OTUs algorithm. **(B)** Associations of rare species discovered by Anets-OTUs in samples comprised clusters 2, 10, and 16. Small components of the network are not included. OTUs are represented by nodes (filled circles) where color indicates different clusters inferred by Markov clustering. The largest clusters are referred as communities. Edges between nodes represent significant associations ($p < 0.001$) between a pair of OTUs. They are labeled by black ovals and have associated bar charts showing the number of OTUs from most abundant taxonomic ranks labeled as G (*Genus*) and O (*Order*). **(C)** Heat map of abundances (in terms of sequencing reads) of associating microbial OTUs (horizontal axis) in three distinct clusters of samples (vertical axis) collected from the human skin. OTUs are grouped according to the microbial communities inferred by Anets-OTUs and sorted by mean abundance; samples are grouped according to clusters inferred by Anets-Samples and sorted by the shared richness. Each cell shows the number of OTU reads. Color of cells in the map shows the number of reads representing the OTUs in the sample: 10 reads or more (dark orange), from 1 to 10 reads (light orange), and not represented by reads (gray). Cluster IDs indicated in **(A,B)** are shown in vertical and horizontal bars of the heat map respectively.

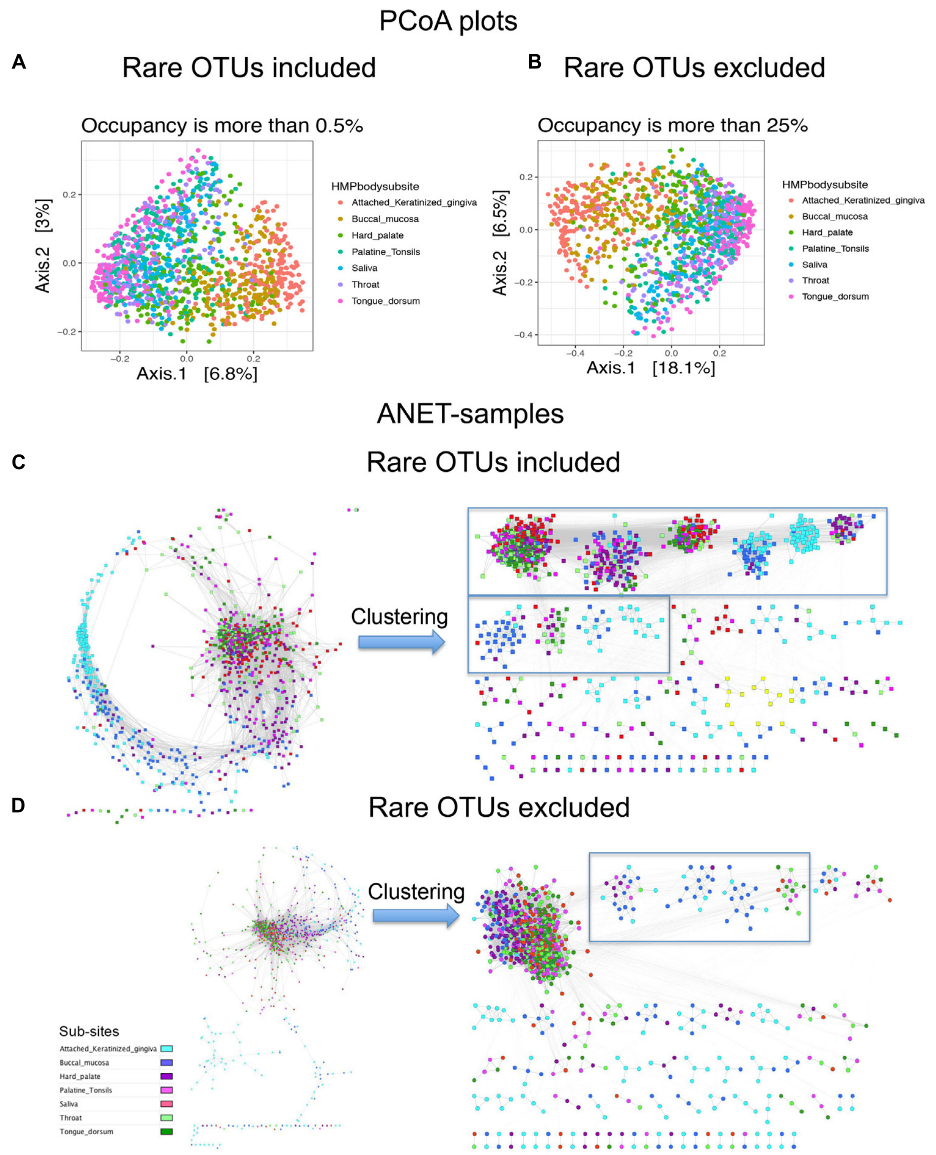


FIGURE 6 | Principal coordinates analysis (PCoA) plots and Anets-Samples for oral samples with or without rare OTUs. **(A)** PCoA plot generated by including rare OTUs. **(B)** PCoA plot generated by excluding rare OTUs. **(C)** Anets-Samples generated by including rare OTUs. **(D)** Anets-Samples generated by excluding rare OTUs. Large clusters (more than 10 samples) are bordered by rectangles.

Validation of the Anets Algorithm

We use 1250 oral samples of HMP to investigate the robustness and limitations of the Anets algorithm, to compare it with other methods and to explore potential biases and confounding factors.

Library Size as Potential Confounding Factor

The Library Size (LS) affects the number of identified rare species and, therefore, may introduce a technical bias in the OTU table if there are significant differences in LSs among studied environments. We explore this affect using known annotations of oral samples by subsites. Specifically, pair-wise comparisons were performed among all the subsites in terms of the library size and then in terms of the number of rare OTUs. We find that log-transformed values of the library size in the oral samples have a normal distribution (Supplementary Figure S4). Significant differences between average values (Wilcoxon test) were observed for 2 out of 15 pair-wise comparisons (Supplementary Figure S5), and only for one comparison, "Tongue dorsum" versus "Hard palate," the difference in LS is also associated with the significantly different number of rare species (Supplementary Table S3). In general, most rare OTUs are the least abundant and the mean number of such OTUs is significantly different in 60% subsite pairs (Supplementary Figure S6 and Supplementary Table S3). When we consider less rare OTUs we find a significant increase in the mean abundance of the OTUs (Supplementary Figure S6) and significant decrease in the % of subsite pairs that are significantly different in terms of the number of rare OTUs, from 60 (occupancy threshold 1%) to 40, 20, and 13% (occupancy threshold 5, 10, and 25%, respectively) (Supplementary Table S3). According to the results, the LS may be a confounding factor in the analysis of rare OTUs, although the different LS doesn't necessary translate to different number of rare species, at least for oral subsites. There is a clear trend for oral subsites to be less different in terms of the number of rare OTUs when we increase the occupancy threshold. This trend, however, doesn't associate with different LSs of the subsites.

Importance of Rare OTUs for Anets-Samples Construction

We further explore how important rare and common taxa for correct grouping of samples. We separated species identified in 1250 oral samples to two groups, rare (occupancy is between 0.5 and 25% samples) and common (occupancy > 25%). Then we generated three OTU tables; comprised of only rare OTUs, rare and common OTUs, and only common OTUs. We find that considering only rare OTUs we reduce the resolution of the Principal coordinates analysis (PCoA) plot (Supplementary Figure S7A). In case of Anets-Samples (Supplementary Figure S7B), we actually increase the resolution and were able to detect a batch effect among oral samples. The effect was probably masked by the presence of common species because we didn't observe the effect if we use OTU table with only common OTUs (**Figure 6D**) or with common and rare OTUs (**Figure 6C**). In spite of the batch effect, the grouping of samples within the large batch (Supplementary Figure S7B) was consistent with the studied oral subsites, although not as evident as for Anets-Samples based on a combined set of rare and common OTUs (**Figure 6D**). The PCoA plots generated for OTU tables by including or excluding the rare OTUs were rather similar (**Figures 6A,B**) suggesting that we will not significantly effect the interpretation of the results by excluding rare species in the PCoA. However, by excluding the rare species when building Anets (**Figure 6D**), we essentially decrease our chance to cluster samples according to subsides (**Figures 6C,D,** right sides) and also decrease the number of associated samples ($p > 0.05$) from 1082 (87%) to 981 (78%). The results demonstrate high sensitivity of the Anets algorithm to signals from both, rare and more abundant, OTUs. The result is not surprising. To build Anets we have to collect additional statistics on co-occurrence of species with the rest and on the shared species richness to establish the pair-wise associations in Anets-Samples and in Anets-OTUs. By excluding some species, either less abundant or more abundant, we loose information important for the analysis and impair the results. Building Anets after filtering common species, however, may allow us to see biases obscured by the presence of common taxa.

Topological Differences Between Networks Generated Using Anets and Unweighted UniFrac Distances

UniFrac is widely used distance metric incorporating phylogenetic information to compare microbial communities. All taxa, common and rare, are included in calculation of the distance. The metric, therefore, may be an alternative way to construct the network of samples by incorporating the phylogenetic signals from rare species. We have compared the network of samples generated by Anets with those based on the Unweighted UniFrac (UUF) distances. The ‘phyloseq’ package (McMurdie and Holmes, 2013) was used to calculate the UUF distance for each pair of oral samples. Two networks were generated with thresholds for the distance to be equal 0.95 and 0.98. We chose these thresholds because we find it difficult to break the UniFrac-based networks into clusters because of low clustering coefficients and high centralization if compared with the Anets-Samples (Supplementary Table S4). We could increase the clustering coefficient and reduce centralization by increasing the distance measure but it also reduced the number of nodes in the UUF network. Using a looser threshold (0.95) we had 1243 nodes that were vastly interconnected by 68284 edges into one large cluster (Supplementary Figure S8). By increasing the distance threshold to 0.98 we generated a network with 868 samples and 6457 edges and a greater clustering coefficient (**Figure 7B**). The generated clusters, however, were not as consistent with the annotation of subsites as in case of Anets-Samples network (**Figure 7A**). Although in general all three networks showed the same trend of separation of subsites ‘keratinized gingiva’ and ‘bunccal mucosa’ from ‘saliva,’ ‘tongue dorsum,’ and ‘throat,’ it was easier to cluster the Anets-based network, and, importantly, many large clusters in the Anets network were enriched with samples originated from the same subsite (**Figure 7A**, right side). The comparison reveals a distinct topology of the Anets network if compared with UUF-based networks and a better association of the topological structure with oral subsites. The more centralized topology of the UUF-based network may be suitable for a global overview of the samples. The Anets-based network may perform better if we want a greater level of detail and more granularity in grouping the samples.

Robustness of the Anets Algorithm

Several different factors including sampling strategy and sample handling, the choice of universal 16S rRNA gene PCR primers, DNA extraction methods, amplification artifacts, such as chimeras, and computational methods employed to produce the OTU table from sequencing reads may contribute to different results in the 16S rRNA gene profiling studies. All of them can affect the number of rare species and the produced Anets. To evaluate the robustness of the algorithm we explore changes in the structure of Anets based on OTU tables constructed by different processing pipeline, by different 16S rRNA gene variable region for sequencing, and by a different subset of oral samples. Namely, we consider three different OTU tables produced for oral samples by two commonly used 16S rRNA amplicon data processing pipelines, MOTHUR (Schloss et al., 2009) and QIIME (Caporaso et al., 2010) that utilize different algorithms to construct the OTU table. The former OTU table was produced by a high quality-filtering MOTHUR pipeline (Schloss et al., 2011) with low overall chimera rate. The formation of the chimeric sequences is a well-known factor contributing

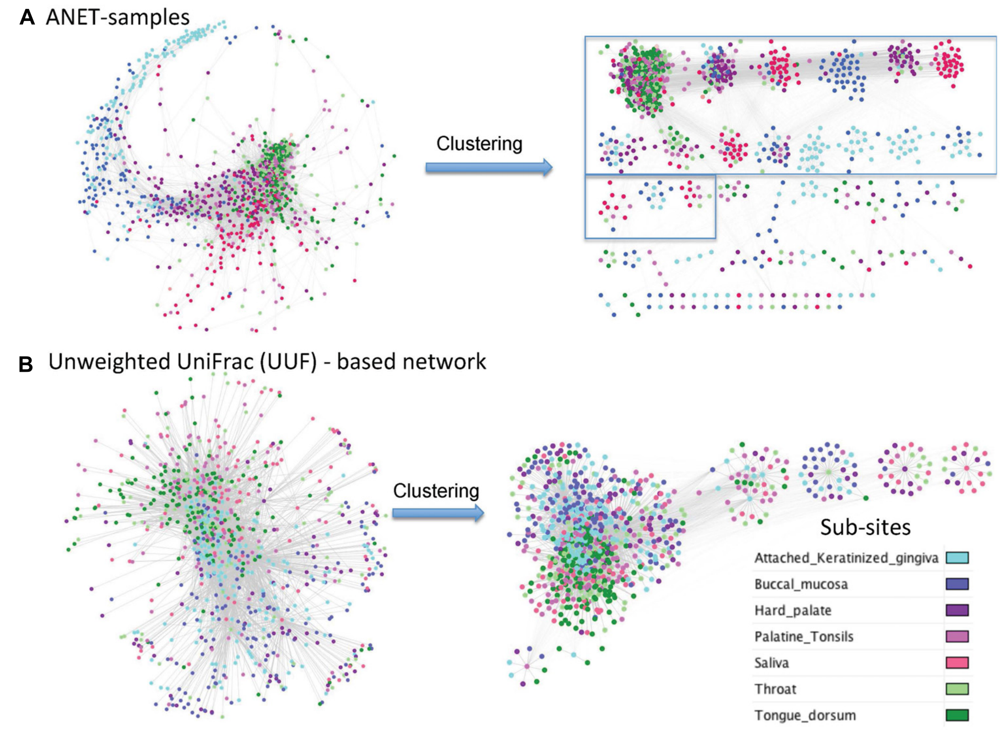


FIGURE 7 | Networks of oral samples and their clustering by the Markov clustering algorithm (MCL) with the same parameters. **(A)** The network was generated using Anets-Samples algorithm. The large clusters (more than 10 samples) are bordered by rectangles. **(B)** The network was generated using Unweighted UniFrac (UUF) distances as measure of pair-wise similarity of the samples (nodes) with the threshold 0.98.

to erroneous OTUs and to overrated species richness (Ashelford et al., 2005). We also compared OTU tables generated by QIIME pipeline from sequencing of 16S rRNA gene variable regions 1–3, referred as HMP v13 (Q), and variable regions 3–5, referred as HMPv35(Q). These three OTU tables were generated for the same subset of 1250 oral samples. In addition, we included an OTU table (QIIME pipeline, v35) produced for a different subset of 1025 oral samples in the comparison. We refer to the table as HMPv35(Q) validation. The tables were downloaded from the NIH Human Microbiome Project websites and were comprised of different number of OTUs, from 8640 OTUs in HMPv13(M) to 26399 OTUs in HMPv35(Q) Validation. Most OTUs (95–97%) in the tables were rare OTUs (found in less than 25% samples). The Anets-Samples was generated for each OTU table and visualized by Cytoscape using the same parameters. Comparison of the produced networks reveals not only their similar statistical characteristics (Supplementary Table S5), but also a similar trend in grouping of samples among subtypes (**Figure 8**). The MOTHUR and QIIME networks, however, were surprisingly different in their ability to separate different subsites (**Figures 8A,B**). The MOTHUR network performed well in separating tongue dorsum and throat from other subsites, but not as good in separating keratinized gingiva and buccal mucosa, while the QIIME v13 network performed better in separating keratinized gingiva and buccal mucosa from other subsites, and not as good for tongue dorsum and throat. The difference persists when we run Anets with different parameters. An interesting symmetrical structure, related to the batch effect, was revealed in the Anets-samples produced for OTU table HMPv35(Q) (**Figure 8C**). The upper part of the network represents samples sequenced by J. Craig Venter Institute (JCVI) and the lower part representing samples sequenced by other sequencing centers. Importantly, each side of the network demonstrated similar grouping of samples into subtypes regardless of the batch affect. The network generated for a different subset of oral samples, HMPv35(Q) Validation, reveal a similar batch effect with separation of samples into subsites within each batch. Based on the results we conclude that the Anets algorithm recover similar groupings of samples from OTU tables produced by two commonly used 16S rRNA amplicon data processing pipelines regardless of the observed batch effects and type of sequencing (v13 or v35) as well as from an OTU table comprised of different samples from the same environments.

DISCUSSION

In this proof of concept study we aimed to demonstrate the use of the Anets-based computational framework for linking associations of rare OTUs to their environment. Results of the study demonstrate that a combination of the Anets-OTUs and Anets-Samples has a potential to serve as a powerful unsupervised methods for discovering relationships and associations of rare species from phylogenetic marker gene

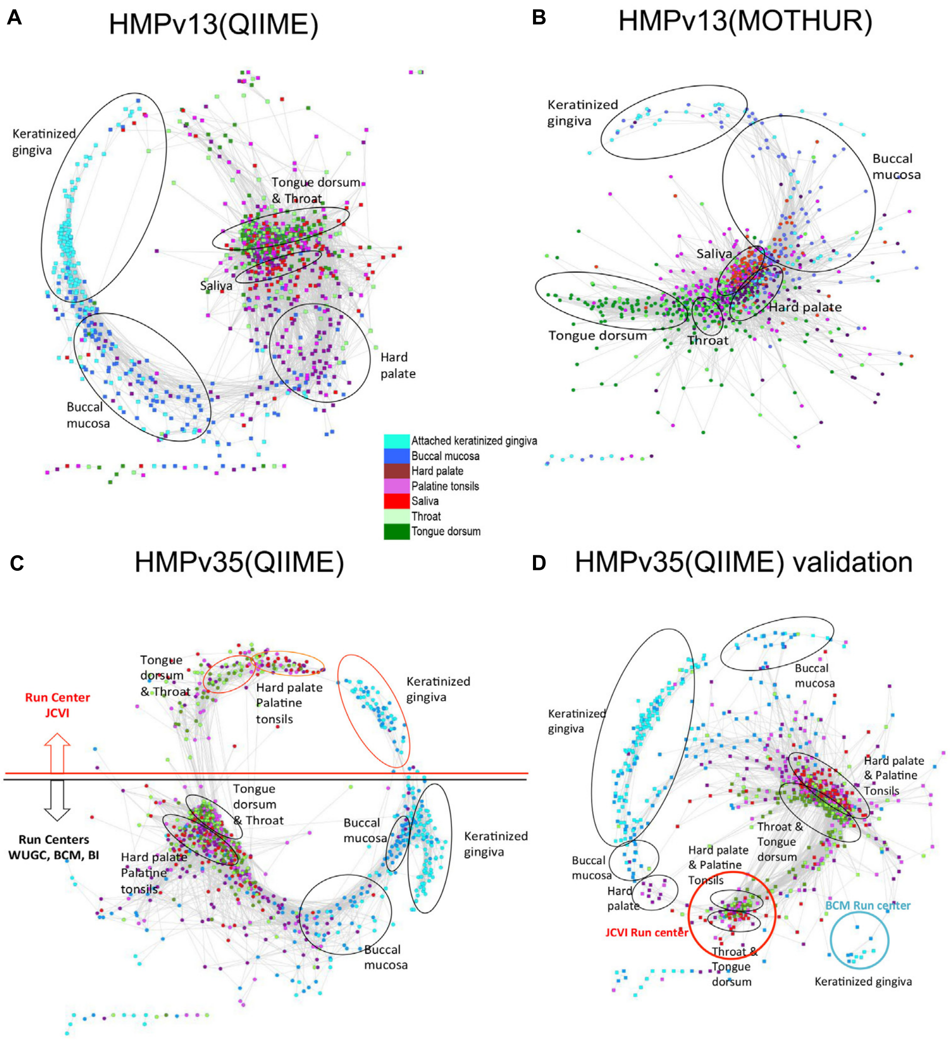


FIGURE 8 | Anets-Samples generated for different OTU tables comprised of oral samples. **(A)** OTU table generated by QIIME pipeline from sequencing of 16S rRNA gene variable regions 1–3. **(B)** OTU table generated by MOTHUR pipeline from sequencing of 16S rRNA gene variable regions 1–3. **(C)** OTU table generated by QIIME pipeline from sequencing of 16S rRNA gene variable regions 3–5. **(D)** OTU table generated by QIIME pipeline from sequencing of 16S rRNA gene variable regions 3–5 of a distinct set of oral samples.

datasets used in microbiome studies. Applying the framework to analyses of microbiomes in *Populus* roots and on Human body sites we were able to reproduce associations of samples in these complex environment and associations of species that were consistent with the existing metadata and the analyses described in the previous literature. In case of Human microbiomes we were able to identify associations of co-dependent rare OTUs and link them to sub-sides of the human body. Similar observations were reported by Ding and Schloss (Ding and Schloss, 2014) using the Dirichlet multinomial mixture models (Holmes et al., 2012).

An important observation from the analysis of *Populus* and Human microbiomes by the approach is a close link between the rare microbial OTUs and specific environmental conditions. To explain the importance of rare putative species for classification of the environments we propose that the high-abundance OTUs are common among sampled environments because the environments have some common conditions stimulating outgrowth of the same putative species. The rare low-abundance OTUs are rare because each of these environments also has some specific conditions or microenvironments. These specific microenvironmental conditions may stimulate the growth species represented by rare OTUs. Although they are rare, they may be crucial for recovering the micro-environmental differences in microbiomes of the environments. It is possible that these rare OTUs, therefore, may be a better computational target for quantification of subtle differences among most variable properties of the environments, and their presence/absence pattern can be used for additional comprehensive classification of samples from the environments. New approaches to 'denoising' sequencing data that avoid collapsing OTUs to higher taxonomic levels or *a priori* OTU similarity thresholds, such as ASVs approach (Callahan, 2017), might also further increase the ability to recover the micro-environmental differences among samples.

Although the results show the importance of rare OTUs in discriminating oral subsites and in revealing batch effects, they don't prove that the rare OTUs are real. Further experimental studies are necessary to provide a direct evidence of their existence. Models of microbial communities where a signal from rare species can be captured and compared with signals from common species would be also helpful to explore rare species and to validate the approach. There are, however, some challenges in developing a realistic model of microbial communities. Available computational tools, such as "SPIEC-EASI" R package (Kurtz et al., 2015) generate a synthetic OTU data using a random selection of species. The randomness contradicts the major assumption of the Anets algorithm that the selection of species in the sample is not random. In addition, the OTU tables simulated by a random selection don't necessary conform to the occupancy–abundance relationship (Gaston, 1996) observed in real settings.

The transformation of OTU table into the OTU presence/absence values for analysis by Anets places some limitations and constraints on the approach. One such constraint is the presence of many common OTUs, such as found in more than ∼75% samples. The loss of abundance data is another limitation. The information can be important for understanding dominant taxa and their interdependencies with each other and members of the rare biosphere. Another important condition for successful application of the approach is the species co-dependence in the studied environments. The condition is important to observe similar co-occurrence profiles for the associated OTUs and to simplify their clustering. Although this assumption is consistent with known metabolic and functional dependences of microbial species in different environments (Jousset et al., 2017), these dependences are not always the major factors that discriminate environments in a particular study.

Further studies are necessary to validate the proposed framework, to extend it by incorporating additional statistical tools, to provide guidelines on setting parameters for the Anets-Samples and Anets-OTUs, to explore different measures of similarity and their cutoffs, and to clarify limitations of the approach. Further work is also necessary to streamline all calculations in a package. At this point, the computations proposed in the framework are implemented by different programs, such as Anets (Karpinets et al., 2012), Cytoscape (Smoot et al., 2011), mcl (Markov clustering) (Van Dongen, 2008), as well as by simple in-house scripts written in R (see "Operating Procedure to generate Anets" in Supplementary Data Sheet S1). Importantly, the Anets program was implemented for a single processor to cope with a data of small scale and complexity. The program will be slow in processing large OTU tables generated by increasingly complex datasets. It is important to increase scalability of the algorithm by parallelizing independent computation steps and by designing efficient representation of the sparse data for better memory management.

We have thus taken the first initial steps in incorporating the "rare biosphere" of microbial community data and linking their contribution to environmental and phenotypic characteristics via the Anets algorithm. More interesting relationships may be found by this approach as the rate of accumulation of microbial data in different environments continues to increase and the cost of sequencing continues to decrease. We believe that the Anets technique holds unexplored potential for an in-depth analysis of the data. The approach is useful to reveal inherent patterns in the data without *a priori* knowledge of factors influencing the microbial communities as well as to visualize the patterns as networks or maps.

MATERIALS AND METHODS

Mock Dataset

The dataset was generated manually to illustrate the ANETs approach, and represents an oversimplified case of two artificial environments populated by eight hypothetical species. The environments were randomly sampled in 12 locations as described in **Figure 3A** in more detail. The major goal of the dataset was to provide an intuitive illustration of the proposed framework.

Populus Root Microbiome Dataset

The dataset was described by Shakya et al. (2013). It includes 84 samples that represent a combined (fungal and bacterial) microbiome in rhizoshpere (46 samples) and endosphere (38 samples) of 23 mature *Populus deltoids* trees growing in Tennessee (11 trees) and North Carolina (12 trees) taken in May (23 rhizosphere samples and 21 endosphere samples) and in September (23 rhizosphere samples and 17 endosphere samples). Bacterial (16S rRNA) and fungal (28S rRNA) genes from the samples were sequenced to estimate the abundance of fungal and bacterial OTUs and their association with plant phenotypic, genotypic, and environmental parameters. We initially explore abundance–occupancy relationships in the dataset using all rhizosphere and endosphere samples of the study (**Figure 2**) and then focused our further analysis on 46 rhizosphere samples. The OTU table for these samples was processed using the Anets tool in two ways: (1) to build the association network of OTUs, Anets-OTUs, and (2) to build the association network of samples, Anets-Samples. The Anets-Samples was generated using the Pearson correlation as the measure of association for each pair of samples and a *p*-value threshold equal 0.01. The Anets-OTUs was generated using OTUs that occurred in 10 or more samples. This threshold was necessary to reduce time and memory used by the Anets program for processing the data. The *p*-value threshold was set to 0.05. Markov clustering (Van Dongen, 2008) with the inflation value 1.8 was used to cluster the networks, and Cytoscape (Smoot et al., 2011) was used to visualize the networks. Soil properties for samples collected near 23 trees were analyzed using hierarchical clustering. All soil parameters were normalized before the clustering using the average value of the parameter and its standard deviation. The hierarchical clustering of soil samples was performed using Pearson correlation as the similarity metric and centroid linkage as the clustering method. The analysis was implemented using the Cluster 3 program (Eisen et al., 1998). The Java Treeview[2] was used to visualize the clusters. The 'vegan' R package (Dixon, 2003), function 'capscale,' was used to calculate variance partitioning the same way as in the initial study (Shakya et al., 2013).

Human Microbiome Dataset

The dataset was downloaded from the HMP website http://www.hmpdacc.org/HMQCP/. The dataset is based on the analysis of 16S rRNA gene variable regions 1–3 (V13) and includes 2910 samples obtained from 18 different body sites of 180 healthy men and women. Each site was represented by 145–190 samples, except the vagina (87–89 samples). The data is described in more detail by the Human Microbiome consortia publications (Project, 2012). The input for the analysis was the OTU table generated by the project from sequencing reads by the QIIME (Quantitative Insights Into Microbial Ecology) software (Caporaso et al., 2010). The table is comprised of 43140 OTUs and 2910 samples. For the cluster enrichment analysis we used publically available sample metadata, sex of the participant and body site.

The downloaded OTU table was processed using the Anets-Samples algorithm to build the association network of samples. The network was generated using the Pearson correlation as the measure of association for each pair of samples and the p-value threshold 0.01. The p-values were calculated using a Monte Carlo simulation approach as described before (Karpinets et al., 2012). The network was visualized using edge-weighted (by p-value) spring embedded layout in Cytoscape (Smoot et al., 2011). The Anets-OTUs was generated for a subset of 314 skin samples selected by the analysis as significantly associated (clusters with IDs 2, 10, and 16 in **Figure 2**). The OTUs table of the samples was used as input for the Anets-OTUs algorithm with the following parameters: the minimum number of samples per OTU is 15, and a p-value threshold is 0.001. The stringent thresholds were important to limit memory use and the processing time for the Anets program. Markov clustering (Van Dongen, 2008) with the inflation value 1.8 was used to cluster the networks, and Cytoscape (Smoot et al., 2011) was used to visualize the networks and the clustering results. An edge-weighted (by p-value) spring embedded layout was used for the network visualization.

Enrichment Analysis

The analysis was used to find samples enriched in each cluster of OTUs in the Anets-OTUs and to find phenotypic or environmental characteristics enriched in each clusters of samples in the Anets-Samples. In both cases the analysis was done using the Fisher's exact test to examine independence of rows and columns in a two-dimensional contingency table generated by the following algorithms.

We identified samples enriched in the cluster of OTUs (Anets-OTUs) by linking each clustered OTU to the sample and finding those samples that have the greatest representation by OTUs within the cluster. We used the fisher.test() function in R to calculate probability that the number of OTUs representing a sample in the cluster is significantly greater than the number expected by randomly selecting OTUs in the cluster from a set of all associated OTUs, regardless of sample of origin. All associated OTUs were found as a set of unique OTUs associated significantly (p-value < 0.05) with at least one other OTU in the Anets-OTUs. We classified the associated OTUs in two ways: if the OTU belongs to the sample or not, and if the OTU belongs to the cluster or not. Using this classification we created the contingency table with the number of the sample's OTUs in the cluster, the number of associated OTUs in the sample, the number of OTUs in the cluster that are not from the sample, and the number of associated OTUs that are not found in the sample. Because we performed several statistical tests simultaneously on the same data set, p-values calculated by the Fisher exact were adjusted using Bonferroni correction.

Specific characteristics (such as soil conditions in the *Populus* rhizosphere dataset or body subsites in the HMP dataset) enriched in the cluster of samples (Anets-Samples) were identified by linking each sample to the characteristics and revealing the characteristics represented by the greatest number of samples within the cluster. We used the Fisher's exact test to calculate probabilities that number of samples representing a characteristic within the cluster is significantly greater than the number expected by randomly selecting samples into the cluster from a set of all associated samples. In this case the background of the comparison was a set of all associated samples; they were classified for each cluster and each characteristic to create the contingency table as (i) representing the environmental/phenotypic characteristic or not and (ii) belonging to the cluster or not.

Generating Networks and Their Statistics for Validation

All datasets for validation were downloaded from the HMP website from the link https://www.hmpdacc.org/hmp/HMMCP/ for 16S rRNA amplicon datasets processed by QQIME and the link https://www.hmpdacc.org/hmp/HMQCP/ for datasets processed by MOTHUR software package using a high stringency approach (Schloss et al., 2011). The 'phyloseq' R package (McMurdie and Holmes, 2013) was used to download the datasets, to create OTU tables for oral samples for comparisons, to filter OTUs by occupancy, to generate the UUF distances (default parameters) and to produce PCoA plots (distance measure was set to 'binary'). The Anets-Samples were generated using Pearson correlation as measure of similarity and setting p-value threshold to 0.05. The networks were loaded into Cytoscape software, visualized using spring embedded layout without edge weighting and clustered using MCL algorithm by a Cytoscape plugin 'clusterMaker2'[3] by setting the inflation value to 2.0. Another Cytoscape plugin 'Network Analyzer'[4] was used to explore topology of the networks and to produce their statistics.

[2] http://sourceforge.net/projects/jtreeview/

[3] http://www.rbvi.ucsf.edu/cytoscape/clusterMaker2/

[4] http://apps.cytoscape.org/apps/networkanalyzer

AUTHOR CONTRIBUTIONS

TK conceived the study. CWS contributed to the preparation, collection, and analysis of the data. JW, AF, CWS, and JZ provided mentoring guidance and advices throughout the study. TK, VG, JW, AF, CWS, and JZ contributed to writing the manuscript.

FUNDING

This work was made possible through support from the Moon Shorts Programs at the University of Texas MD Anderson Cancer Center and from the Genomic Science Program, United States Department of Energy, Office of Science, Biological and Environmental Research, as part of the Plant Microbe Interfaces Scientific Focus Area and the BioEnergy Science Center (BESC). Oak Ridge National Laboratory is managed by UT-Battelle LLC, for the United States Department of Energy under contract DE-AC05-00OR22725. The submitted manuscript has been authored by a contractor of the United States Government under contract DE-AC05-00OR22725.

ACKNOWLEDGMENTS

We would like to thank Dr. Michael Robeson and Dr. Migun Shakya for their comments, editing the manuscript, and for helpful discussions. We thank Dr. Migun Shakya for providing us a script to calculate variance partitioning the same way as in his original study. We also thank the reviewers of the manuscript for their helpful comments and analysis suggestions that greatly improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00297/full#supplementary-material

REFERENCES

Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., et al. (2013). The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* 27, 1012–1022. doi: 10.1096/fj.12-220806fj.12-220806

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71, 7724–7736. doi: 10.1128/aem.71.12.7724-7736.2005

Bascompte, J., Jordano, P., Melian, C. J., and Olesen, J. M. (2003). The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9383–9387. doi: 10.1073/pnas.16335761001633576100

Callahan, B. J. (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, 581–583. doi: 10.1038/Nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303nmeth.f.303

Coveley, S., Elshahed, M. S., and Youssef, N. H. (2015). Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (Zodletone spring. OK, USA). *PeerJ* 3:e1182. doi: 10.7717/peerj.1182

Cripps, C. L. (1997). The genus Inocybe in Montana aspen stands. *Mycologia* 89, 670–688. doi: 10.2307/3761005

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

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14, 927–930. doi: 10.1111/j.1654-1103.2003.tb02228.x

Edgar, R. C., and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31, 3476–3482. doi: 10.1093/bioinformatics/btv401

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868. doi: 10.1073/pnas.95.25.14863

Fang, H., Huang, C., Zhao, H., and Deng, M. (2017). gCoda: conditional dependence network inference for compositional data. *J. Comput. Biol.* 24, 699–708. doi: 10.1089/cmb.2017.0054

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.1002606PCOMPBIOL-D-12-00158

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

Gaston, K. J. (1996). The multiple forms of the interspecific abundance-distribution relationship. *OIKOS* 76, 211–220. doi: 10.2307/3546192

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.0030126PONE-D-11-15801

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234nature11234

James, A., Pitchford, J. W., and Plank, M. J. (2012). Disentangling nestedness from models of ecological complexity. *Nature* 487, 227–230. doi: 10.1038/nature11214nature11214

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., et al. (2017). Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.* 11, 853–862. doi: 10.1038/ismej.2016.174

Karpinets, T. V., Park, B. H., and Uberbacher, E. C. (2012). Analyzing large biological datasets with association networks. *Nucleic Acids Res.* 40:e131. doi: 10.1093/nar/gks403

Kurtz, Z. D., Muller, 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

Lynch, M. D., and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* 13, 217–229. doi: 10.1038/nrmicro3400

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217PONE-D-12-31789

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

Mi, X., Swenson, N. G., Valencia, R., Kress, W. J., Erickson, D. L., Perez, A. J., et al. (2012). The contribution of rare species to community phylogenetic diversity across a global network of forest plots. *Am. Nat.* 180, E17–E30. doi: 10.1086/665999

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pedros-Alio, C. (2012). The rare bacterial biosphere. *Ann. Rev. Mar. Sci.* 4, 449–466. doi: 10.1146/annurev-marine-120710-100948

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl. Environ. Microbiol.* 79, 6593–6603. doi: 10.1128/Aem.00342-13

Project, T. H. M. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209nature11209

Rosindell, J., Hubbell, S. P., and Etienne, R. S. (2011). The unified neutral theory of biodiversity and biogeography at age ten. *Trends Ecol. Evol.* 26, 340–348. doi: 10.1016/j.tree.2011.03.024S0169-5347(11)00094-2

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6:e27310. doi: 10.1371/journal.pone.0027310

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 77, 3219–3226. doi: 10.1128/Aem.02810-10

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.01541-09AEM.01541-09

Shakya, M., Gottel, N., Castro, H., Yang, Z. K., Gunter, L., Labbe, J., et al. (2013). A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature populus deltoides trees. *PLoS One* 8:e76382. doi: 10.1371/journal.pone.0076382PONE-D-13-28933

Sharon, I., Kertesz, M., Hug, L. A., Pushkarev, D., Blauwkamp, T. A., Castelle, C. J., et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* 25, 534–543. doi: 10.1101/gr.183012.114

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675btq675

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Suweis, S., Simini, F., Banavar, J. R., and Maritan, A. (2013). Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* 500, 449–452. doi: 10.1038/nature12438nature12438

Tsilimigras, M. C., and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330–335. doi: 10.1016/j.annepidem.2016.03.002

Unterseher, M., Jumpponen, A., Opik, M., Tedersoo, L., Moora, M., Dormann, C. F., et al. (2011). Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology. *Mol. Ecol.* 20, 275–285. doi: 10.1111/j.1365-294X.2010.04948.x

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi: 10.1137/040608635

Youssef, N. H., Couger, M. B., and Elshahed, M. S. (2010). Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, USA): the role of the rare biosphere. *PLoS One* 5:e12414. doi: 10.1371/journal.pone.0012414e12414

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.

ORIGINAL RESEARCH
published: 14 December 2017
doi: 10.3389/fmicb.2017.02507

Modeling Metabolic Interactions in a Consortium of the Infant Gut Microbiome

*Francisco Pinto, Daniel A. Medina, José R. Pérez-Correa and Daniel Garrido**

Department of Chemical and Bioprocess Engineering, School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

The gut microbiome is a complex microbial community that has a significant influence on the host. Microbial interactions in the gut are mediated by dietary substrates, especially complex polysaccharides. In this environment, breakdown products from larger carbohydrates and short chain fatty acids are commonly shared among gut microbes. Understanding the forces that guide microbiome development and composition is important to determine its role in health and in the intervention of the gut microbiome as a therapeutic tool. Recently, modeling approaches such as genome-scale models and time-series analyses have been useful to predict microbial interactions. In this study, a bottom-up approach was followed to develop a mathematical model based on microbial growth equations that incorporate metabolic sharing and inhibition. The model was developed using experimental *in vitro* data from a system comprising four microorganisms of the infant gut microbiome (*Bifidobacterium longum* subsp. *infantis*, *Lactobacillus acidophilus, Escherichia coli,* and *Bacteroides vulgatus*), one substrate (fructooligosaccharides, FOS), and evaluating two metabolic products (acetate and lactate). After parameter optimization, the model accurately predicted bacterial abundance in co-cultures from mono-culture data. In addition, a good correlation was observed between the experimental data with predicted FOS consumption and acid production. *B. infantis* and *L. acidophilus* were dominant under these conditions. Further model validation included cultures with the four-species in a bioreactor using FOS. The model was able to predict the predominance of the two aforementioned species, as well as depletion of acetate and lactate. Finally, the model was tested for parameter identifiability and sensitivity. These results suggest that variations in microbial abundance and activities in the infant gut were mainly explained by metabolic interactions, and could be properly modeled using Monod kinetics with metabolic interactions. The model could be scaled to include data from larger consortia, or be applied to microbial communities where sharing metabolic resources is important in shaping bacterial abundance. Moreover, the model could be useful in designing microbial consortia with desired properties such as higher acid production.

Keywords: metabolic interaction, gut microbiome diet, prebiotics, mathematical modeling, fructooligosaccharides (FOS)

OPEN ACCESS

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Angela M. Zivkovic,
University of California, Davis,
United States
Emanuele Bosi,
University of Florence, Italy

****Correspondence:***
Daniel Garrido
dgarridoc@ing.puc.cl

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

Received: *07 September 2017*
Accepted: *01 December 2017*
Published: *14 December 2017*

Citation:
Pinto F, Medina DA, Pérez-Correa JR and Garrido D (2017) Modeling Metabolic Interactions in a Consortium of the Infant Gut Microbiome. Front. Microbiol. 8:2507. doi: 10.3389/fmicb.2017.02507

INTRODUCTION

The human colonic microbiome is a complex microbial community that has a significant impact on host health. This is a diverse community that reaches high cell densities and includes four dominant phyla (*Bacteroidetes*, *Firmicutes*, *Actinobacteria* and *Proteobacteria*) (Lozupone et al., 2012; Qin et al., 2013). The gut microbiome coexists with the host and deploys important functions that impact host metabolism and gut physiology (Rajilić-Stojanović and de Vos, 2014). Even though its composition is variable among people (Goodrich et al., 2014), the functions these microorganisms performed are basically conserved (Lozupone et al., 2012; Qin et al., 2013). In certain cases, imbalances in the composition of the microbiome are a contributing factor to the onset of inflammatory bowel diseases as well as autoimmune and metabolic (Greenblum et al., 2012; Sevelsted et al., 2015; Marchesi et al., 2016; Tamburini et al., 2016; Cox et al., 2017). How the microbiome assembles in the first months of life appears to be important later in life (Tamburini et al., 2016). One important factor shaping the early microbiome is the type of feeding (Qin et al., 2013; Rajilić-Stojanović and de Vos, 2014). Human breast milk contains large amounts of oligosaccharides (HMO), which are selectively utilized by beneficial gut microbes. *Bifidobacterium* species such as *B. longum* subsp. *infantis* display multiple adaptations to utilize these substrates (Thomson et al., 2017). Lactobacilli are also abundant in the infant gut microbiome (Bäckhed et al., 2015). In contrast, formula-fed infants have a distinct microbiome composition, not dominated by *Bifidobacterium* and with a higher representation of members of *Bacteroides* (*B. fragilis*, *B. vulgatus*) and *Enterobacteria* (*Escherichia coli*, *Klebsiella* spp.) (Bäckhed et al., 2015). The activity of these microbes results in high amounts of acetate and lactate in infant feces, resulting in an acidic pH (Cinquin et al., 2004; Tamburini et al., 2016).

Microbial interactions are important for the assembly and functioning of the gut microbiome. Dominant ecological interactions found in the gut microbiome are competition and cooperation (Faust and Raes, 2012). These interactions broadly represent the sum of all physical, chemical and microbiological activities that microorganisms exert upon others (Roume et al., 2015; Vogt et al., 2015; Hecht et al., 2016; Rakoff-Nahoum et al., 2016). Considering that diet is a major driver guiding gut microbiome composition, microbial interactions are influenced by dietary compounds (Cameron et al., 2014; Medina et al., 2017; Tuncil et al., 2017). Cross-feeding of fermentation breakdown products of the microbiome appears to be common among gut species (Rogowski et al., 2015). This has been shown for example in the utilization of mucin and sialylated milk oligosaccharides between *B. bifidum* and *B. breve* (Egan et al., 2014a,b), or during fructan consumption between bifidobacteria and butyrate- producing bacteria (Moens et al., 2016). Cross-feeding is also observed when metabolic end products from one microorganism, such as amino acids or short chain fatty acids (SCFA), are used by another microorganism (Egan et al., 2014a; Moens et al., 2016). For example, lactate and acetate are end products of lactic acid bacteria, which could be utilized by butyrate-producing bacteria such as *Faecalibacterium prausnitzii* and *Eubacterium rectale* (Louis and Flint, 2017).

Modeling-based approaches have been recently developed to study and predict the composition and interactions in the gut microbiome (Magnúsdóttir et al., 2016). These include ecological-statistic models, genome-scale metabolic reconstructions (GSM) and ordinary differential equation (ODE)-based kinetic models (Trosvik et al., 2010a; Kettle et al., 2015). A Generalized Additive Model (GAM) (Hastie and Tibshirani, 1990) consists of a statistic regression technique that has been used in time-series analysis of ecological data to characterize and estimate cross-feeding and competition between microorganisms. GAMs do not need any assumption about functional relationships in the group for its formulation. However, they could be affected by overfitting when many parameters are needed for matching the data (Wood, 2008; Trosvik et al., 2010b). GAMs usually require a post cross-validation process to curate the model (Ward, 2014). After proper calibration and validation, these models provide accurate predictions by interpolation (Trosvik et al., 2010b).

Lately, GSMs have been successfully applied to explore microbial interactions among gut microbes (Magnúsdóttir et al., 2016). They require an extensive database for reconstruction, editing and gap-filling of full metabolic pathways (Thiele et al., 2014). Several techniques based on orthology, topology and stoichiometry of biological reactions facilitate the draft design and curation process (Thiele and Palsson, 2010). Characteristic features of the species to be reconstructed must be first identified (Kanehisa, 2006). After curation and defining specific environments and constraints, microbial interactions can be obtained for a few species (Thiele et al., 2013).

Recently, a kinetic model constructed from experimental data of gut microbes in a bioreactor was presented, aimed to model the dynamic behavior of the gut microbiome (Kettle et al., 2015). The analysis required a metabolic pathway input and a matrix describing the compounds produced during the fermentation, to generate an ODE system for simulation of microbiome abundance (Walker et al., 2011). Here, microbiome complexity was simplified assigning gut microbes to ten bacterial functional groups (BFGs), based on metabolic properties such as similar breakdown of complex substrates or similar SCFA production or consumption patterns (Kettle et al., 2015). The model showed a good fit with experimental data, which corresponded to a continuous flow bioreactor inoculated with human fecal microbiota.

In order to help understanding the forces dominating gut microbiome structure and composition, here we developed and assessed a mathematical model based on microbial growth equations, taking into account metabolic interactions among bacteria. We focused on the interactions of four gut microbes, *Bifidobacterium longum* subsp. *infantis*, *Lactobacillus acidophilus*, *Bacteroides vulgatus* and *Escherichia coli,* during their growth *in vitro* using fructooligosaccharides (FOS) as substrate. FOS is a well studied prebiotic with degree of polymerization of fructose of 3–6 units (Roberfroid et al., 2010). Experimental data was obtained from co-culture experiments, which were used later to construct and calibrate the model, including the impact of metabolic inhibition or stimulation on bacterial growth. The model was finally validated using additional experimental data of the consortium of the four species on FOS using a biological reactor.

MATERIALS AND METHODS

Microorganisms and Media

Microorganisms used in this study were obtained from the UC Davis, Department of Viticulture and Enology Culture Collection (*L. acidophilus* ATCC 4356, *B. infantis* ATCC 15697, *Escherichia coli* K12), and the American Type Culture Collection (*Bacteroides vulgatus* ATCC 8482; Manassas, VA, United States). Bacteria were, respectively, cultured at 37°C for 24 h in de Man–Rogosa–Sharp (MRS), MRS supplemented with 0.05% L-cysteine-HCl (Loba Chemie, India), LB broth, or Reinforced Clostridium Medium (Becton-Dickinson) supplemented with 1 g/L L-cysteine. All bacteria excepting *E. coli* were routinely grown under anaerobic conditions in an anaerobic jar (Anaerocult, Merck, Germany) with anaerobic packs (Gaspak EM, Becton Dickinson). All media were pre-reduced in an anaerobic jar overnight before inoculation, and prior to each assay bacteria were sub-cultured twice.

Co-culture Batch Experiments

Combinations of *L. acidophilus* (La), *E. coli* (Ec), *B. vulgatus* (Bv) and *B. infantis* (Bi) were prepared in co-culture experiments. Culture media used was a modified version of previously described ZMB (Zhang et al., 2009), which was supplemented with hemin (0.01 g/L, Sigma–Aldrich, St. Louis, MO, United States) and L-cysteine-HCl (0.5 g/L, Sigma–Aldrich, St. Louis, MO, United States). Single amino acid groups in ZMB were replaced by Bacto-Tryptone (at 28 g/L). Carbon sources used were either lactose (10 g/L; Lyngby, Denmark) or FOS (10 g/L; Raftilose Synergy 1, Orafti, Malvern, PA, United States) as carbon source. Single cultures of *B. infantis* (Bi), *B. vulgatus* (Bv), *E. coli* (Ec) and *L. acidophilus* (La); and co-cultures BiBv, BiEc, BiLa, BvEc, BvLa and EcLa were prepared. An experiment with all bacteria (All) and a negative control with no bacteria were included. Fresh overnight cultures of each microorganism were washed in sterile mZMB, and 1 mL of each overnight culture was used to inoculate 10 mL of mZMB containing FOS. This experiment was performed in duplicate. Volumes of 200 μL of inoculated mZMB were placed in 96 well sterile microplates, covered with 30 μL of sterile mineral oil, and incubated in anaerobic jars at 37°C for either 24, 48, or 72 h. In parallel, growth was monitored every 12 h in a microplate reader (Tecan Infinite M200 PRO, Switzerland). Samples were recovered from each microplate and centrifuged at 12000 × *g* for 2 min. Pellets and supernatants were stored at −20°C until use.

Quantification of Bacterial Abundance by qPCR

Total DNA from each sample was purified using the UltraClean® Microbial DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, CA, United States), following manufacturer instructions and using a Disruptor Genie (Scientific Industries, Inc., Bohemia, NY, United States). Extracted DNA was quantified using a NanoQuant Plate in the Tecan Infinite M200 PRO plate reader, and diluted to 1 ng/μL to be used in qPCR reactions. For qPCR we used 0.2 μM of the following primers: for Bv, *Bacteroidetes* primer F (5′-GGTGTCGGCTTAAGTGCCAT-3′) and *Bacteroidetes* primer R (5′-CGGACGTAAGGGCCG TGC-3′); for Bi, Blon_0883F (5′-AGTTCGGCTCCAAAGAC CTG-3′) and Blon_0883R (5′-CATGCCTCGATACGGTCGAA), targeting an ABC solute binding protein; for Ec, Eco1457F (5′-CATTGACGTTACCCGCAGAAGAAG) and Eco1652R (5′-CTCTACGAGACTCAAGCTTGC-3′) (Kassinen et al., 2004); and for La, LACTO_F (5′-TGGAAACAGRTGCTAATACCG-3′) and LACTO_R (5′-GTCCATTGTGGAAGATTCCC-3′) (Bartosch et al., 2004). qPCR reactions were performed using the qPCR PowerUp SYBR Green Master Mix in MicroAmp Fast Optical plates (Applied Biosystems, United States), and using a StepOnePlus Real-Time PCR System (Applied Biosystems, United States). Reactions were carried out for 2 min at 50°C, 2 min at 95°C and 40 cycles of 3 s at 95°C and 30 s at 62°C. Absolute quantification was performed including a standard curve using DNA from a pure culture of each species, with dilutions starting from 1 ng/μL to 0.1 pg/μL. To convert bacterial DNA concentrations into cell genome numbers, the following equation was used (equation 1).

$$\text{Cell copies/mL} = \frac{\text{Avogadro N}^{\circ}\ (1/\text{mol}) \cdot \text{DNA quantity (g/mL)} \cdot \text{Genome 16S copy number}}{\text{Genome size (pb)} \cdot 660(\frac{\text{g}}{\text{mol}})}$$

Batch Bioreactor Culturing

Four independent batch co-culture experiments were performed in a 250 mL bioreactor (Mini-bio Applikon Biotechnology, Netherlands), using mZMB as culture media supplemented with FOS at 1%. In these experiments, the four microorganisms (Bi-La-Ec-Bv) were inoculated at an initial OD_{630} of 0.05. The bioreactor has two six-bladed Rushton turbines and operated at 100 rpm. The temperature was set at 37°C and the pH was maintained at 5.5 with automatic injection of 3N HCl and 3N NaOH. The dissolved oxygen concentration was set at 1 ppm by purging N_2 (99.99% grade) before inoculation and during the lag phase. The foam level was controlled adding 100 μL antifoam in the inoculum (Polydimethylsiloxane base, Winkler, Chile). Two milliliter from the bioreactor were obtained every 2 h and centrifuged at 4000 × *g* for 5 min. Supernatants were stored at −20°C for carbohydrate and SCFA quantification. Pellets were stored for DNA extraction, quantified and diluted to 10 ng/μL for qPCR assays as described above in an AriaMx Realtime PCR System (Agilent Technologies, Santa Clara, CA, United States).

Sample Analysis

Total carbohydrate quantification was performed using the phenol-sulfuric acid method (Tuomivaara et al., 2015). Acetate and lactate were quantified by HPLC using an Aminex HPX-87H ion exchange carbohydrate-organic acid column (Bio-Rad, United States) at 35°C with a flow rate of 0.450 mL/min (H_2SO_4 5 mM, mobile phase) on a LaChrom L-700 HPLC system (Hitachi, Japan), equipped with a Diode Array and a Refractive Index detectors as described previously (Mendoza et al., 2017).

Model Development

The equations used in the model are described in the Model development section in Supplementary Material. The model, the parameter identifiability and sensitivity analysis codes are also presented in Supplementary Material. As input for the determination of the parameters, mono-culture and paired co-culture abundance data are required, in addition to an estimation of acetate and lactate produced and carbohydrate consumed under these conditions. To simplify the analysis, some assumptions were taken into account: (a) an inhibition term was added to Monod kinetics (Model development, Supplementary Material); (b) a microorganism will prefer the consumption of the main carbon sources (glucose, lactose), over other intermediates produced during the fermentation; (c) the ability of a microorganism to produce or consume an intermediate was determined from its metabolic pathway and the literature, and later confirmed experimentally in mono-cultures.

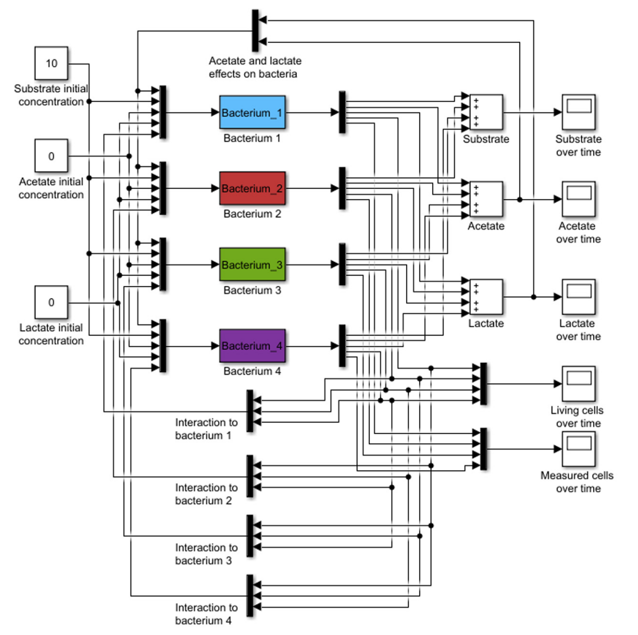

Acetate and lactate effects on bacteria
10
Substrate initial concentration
0
Acetate initial concentration
0
Lactate initial concentration
Bacterium_1
Bacterium 1
Bacterium_2
Bacterium 2
Bacterium_3
Bacterium 3
Bacterium_4
Bacterium 4
Substrate
Substrate over time
Acetate
Acetate over time
Lactate
Lactate over time
Living cells over time
Measured cells over time
Interaction to bacterium 1
Interaction to bacterium 2
Interaction to bacterium 3
Interaction to bacterium 4

FIGURE 1 | Model general representation. Initial substrate and product concentrations and lag phase are used as input (black bars). Microbial growth, consumption, and acid production are considered to interact with other bacteria. Final outputs are observed substrate, acids, and biomass.

RESULTS

Model Description

In this work a kinetic black-box model was developed, aimed to predict the abundance of a bacterial population, substrate consumption and SCFA production, based on mono and co-culture data (**Figure 1**). The model is based on microbial growth equations, but it also considers the metabolic influence of one microorganism on another. This could be considered as a feedback control mechanism (**Figure 1**).

Parameter Settings in Mono-culture

For single microorganisms, the general model consisted of 5 ODEs (Equations 2, 4, 5, and 6 in Supplementary Material), 17 parameters and constitutive Monod-like inhibition equations (Sacher et al., 2011). Mono-culture parameters (**Table 1**) were set as described in the Parameter fitting section in the Supplementary Material. 96 well-plates mono-cultures of Bi, Bv, Ec and La were prepared, in a semi-synthetic media (mZMB) and using FOS as the sole carbon source. Bacterial abundance, FOS consumption and acetate and lactate produced were measured to fit model parameters. An average of eight parameters were set for each bacterium (**Table 1**), which were found by the optimization task. The calculated error in the assay is shown in Supplementary Table S1. For any microorganism and under all conditions, parameter K_s (half-velocity constant) appeared insensitive.

Paired Co-culture and Parameter Fitting

The model was later expanded to include the metabolic interaction between two microorganisms. This model consists of 7 ODEs, 17 parameters per bacteria and two interaction parameters per co-culture. Every parameter not calibrated in mono-culture was set in this step. In order to fit the co-culture parameters, all paired combinations of microorganisms were cultured in FOS and analyzed as described above. **Figure 2A** shows the percentage of change in abundance for all six paired combinations, determined experimentally. As a comparison, **Figure 2B** shows these percentage changes according to the fitted models. Most of the times, the model was able to predict well the changes in abundance in all co-cultures. Experimentally, initial Ec cell numbers were higher than the other microorganisms. However, during growth Bi and La recovered in part their levels compared to Ec (**Figures 2A,B**). Co-culture data allowed the prediction of Bv predominance over La and Bi during growth on FOS, which was also observed experimentally.

Figures 3A,B compares the experimental consumption of FOS by the co-cultures with the values simulated with the fitted model. Most experimental and simulated combinations showed total carbohydrate depletion between 24 and 48 h. In general the model indicated a faster consumption compared to experimental data. One important exception was the BvLa paired co-culture, in which not all of the carbohydrate was consumed. This behavior was not captured by the model, which assumed that since both bacteria reached 100% consumption in single culture, the same rule should apply to their combination.

Figure 4A shows the concentration of acetate produced over time. In certain cases the model predicted the experimental

TABLE 1 | Parameters found via scatter search in mono-culture and then used in co-culture optimization.

Parameter description	Unit	*B. infantis*	*B. vulgatus*	*E. coli*	*L. acidophilus*
K_d	h^{-1}	0.0139	0.0045*	0.0001*	0.0461
K_s	g	0.0412*	0.0203*	0.0190*	0.0027*
μ_{max}	h^{-1}	0.3344	0.3574	0.5063	0.4282
Y_{xs}	$\frac{g_{biomass}}{g_{substrates}}$	0.4268*	0.0054	0.8812	0.1561
I_a	g	29.9727	3.8514	3.4468	0.1723
I_l	g	14.8256	48.8576*	8.4569	17.4298
M_s	$\frac{g_{substrates}}{g_{biomass}} {}^{*}h^{-1}$	0.0003	0.0001	0.0055	0.0001
Y_{ax}	$\frac{g_{biomass}}{g_{acetate}}$	2.0165	5.1489*	4.8884	1*
Y_{lx}	$\frac{g_{biomass}}{g_{lactate}}$	10.2774	10.5967	1*	1.5487*
K_{sa}	g	0.0874*	0.0231	8.6831	1*
K_{sl}	g	0.0615	6.2647	0*	0.3547
β_{maxA}	h^{-1}	0.4687	0.1322	0.284525*	0*
β_{maxL}	h^{-1}	0.1497	0.0006*	0*	0.0011*
μ_{maxA}	h^{-1}	0*	0*	0.1139	0*
μ_{maxL}	h^{-1}	0*	0.0242	0*	0.0074*
Y_{xA}	$\frac{g_{biomass}}{g_{acetate}}$	1*	1*	0.3076	1*
Y_{xL}	$\frac{g_{biomass}}{g_{lactate}}$	1*	0.0247	1*	1*

Set parameters are indicated by (*).

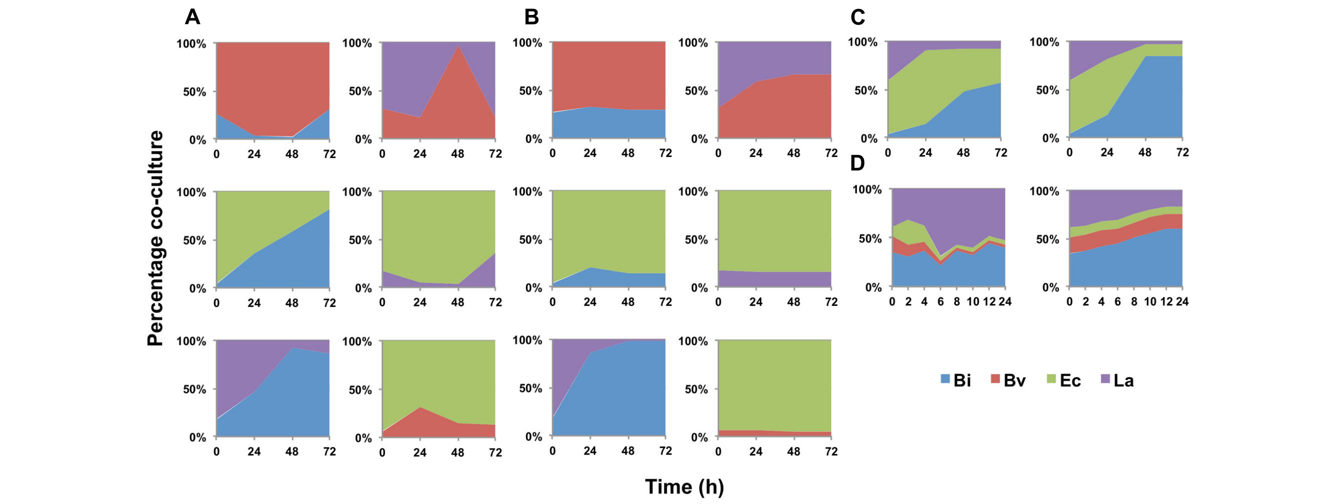


FIGURE 2 | Changes in bacterial population during growth on FOS, expressed as percentage of the co-culture in time. **(A)** experimental data of co-cultures; **(B)** model estimation of abundance in co-cultures; **(C)** abundance of the four-species co-culture in microplates, experimental (Left) and estimated by the model (Right); **(D)** abundance in the four-species co-culture in the bioreactor during growth on FOS, experimental (Left) and estimated by the model (Right).

behavior of acetate production. Bi combinations displayed larger acetate amounts compared to other co-cultures, and in certain cases the model predicted higher values than what was observed. Interestingly, the model predicted that acetate production in co-culture BiEc will have a peak and later decrease. This was also observed experimentally, but at a different time and different intensity (**Figure 4A**). These results indicate that Bi growth is an important parameter for sensitivity assays.

Figure 5A displays the concentration of lactate in co-cultures. A good agreement between observed and predicted data was obtained in co-cultures BiLa, BiEc and LaEc. Combinations BiBv and BvLa were predicted to produce lactate because of Bi and La activities; however, lactate amounts were negligible and not reproduced well by the model. In addition, BvEc co-culture showed production of lactate, but the model assumptions and structure did not consider this situation. The error calculated (equation 10 in Supplementary Material) for the parameter fitting process is shown in Supplementary Table S1.

The parameters determined in paired co-cultures are shown in **Table 1**. The interaction parameters in **Table 2** indicate the influence of one microorganism on another's growth rate. A negative value indicates that one microorganism favors another's growth, while a positive term indicates inhibition.

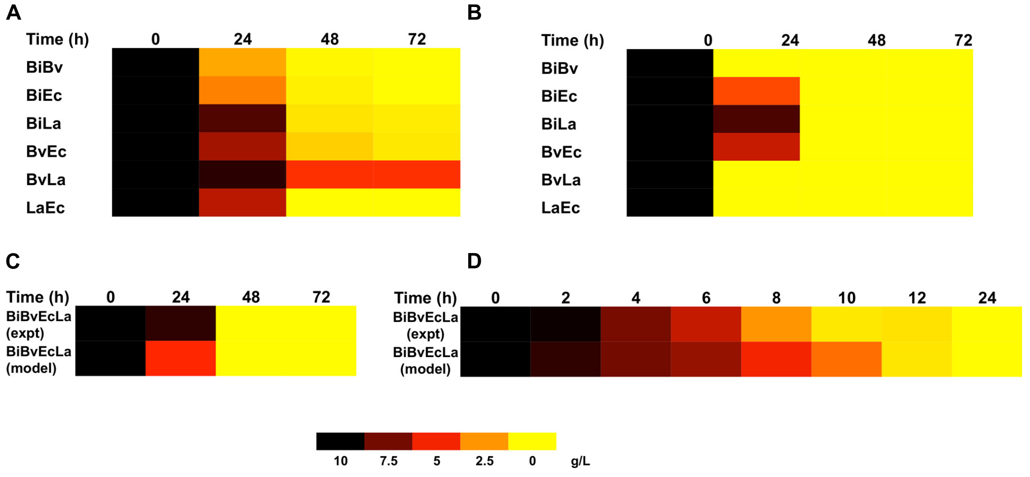


FIGURE 3 | Heat map representing FOS concentration in co-cultures and its prediction by the model. **(A)** FOS concentration in paired co-cultures. **(B)** Model estimation of FOS concentration in paired co-cultures; **(C)** experimental and predicted FOS concentration in the four-species co-culture in microplates; **(D)** experimental and predicted FOS concentration in the four-species co-culture in the bioreactor.

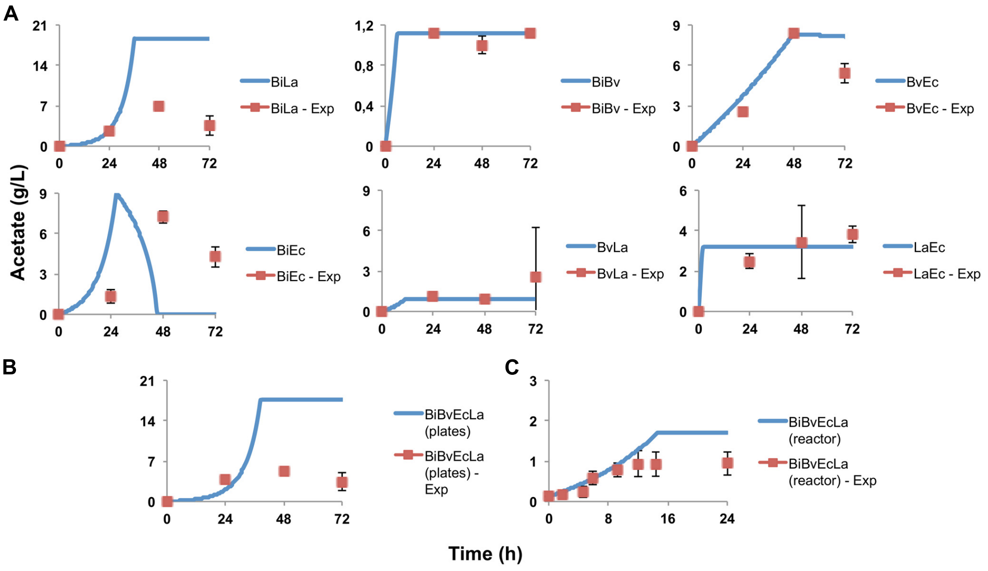


FIGURE 4 | Acetate production and estimation by the model. **(A)** Experimental data in paired co-cultures, compared to values predicted by the model; **(B)** experimental and predicted acetate values of the consortium in microplate assay; **(C)** experimental and predicted acetate values of the consortium in the bioreactor.

Values near 0 suggest a greater interaction effect, while values near the limit indicate there is no effect on the other bacteria. A strong inhibition was found from Ec to Bv, and in general the effects observed were positive or neutral.

Model Validation Using Bacterial Consortia

Finally, the model was validated using independent experimental data from co-culture of the four microorganisms using FOS as the sole carbon source. The experiment was set in microplates and analyzed as discussed above. To test the validity of the model in another set-up, the consortium was additionally cultured on FOS in a 250 mL pH/oxygen controlled stirred bioreactor. This batch system offers a much more controlled and reproducible anaerobic environment, which also provides much faster growth compared to microplates.

Figure 2C shows percentage abundance data obtained for each member of the consortium in microplate assays. The initial levels of Bv were much lower compared to the other three microorganisms. Interestingly, the amounts of La, Ec and Bi in the well-plates cultures were closely predicted by the model. Under these conditions, Bi dominated the co-culture using FOS, followed by La. A good prediction was also observed for the total carbohydrate concentration in spent media (**Figure 3C**). Finally,

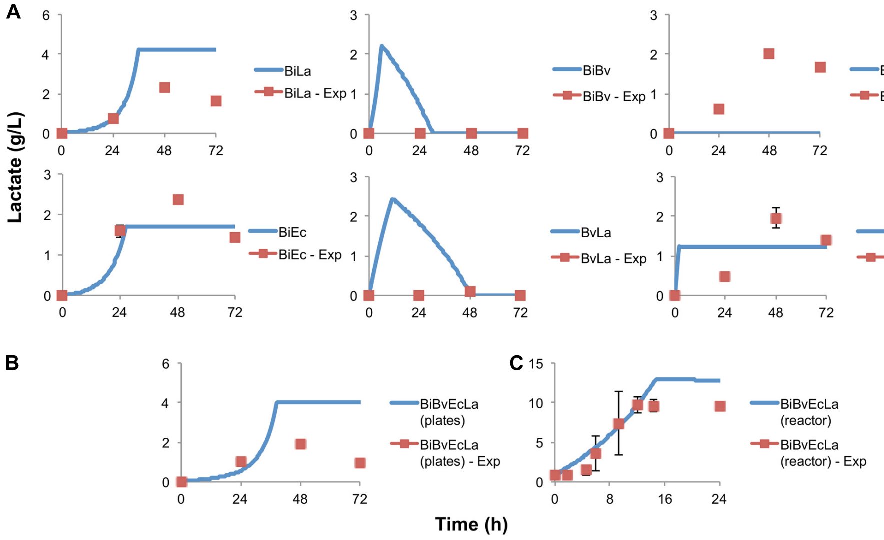


FIGURE 5 | Lactate production and estimation by the model. **(A)** Experimental data in paired co-cultures, compared to values predicted by the model; **(B)** experimental and predicted lactate values of the consortium in microplate assay; **(C)** experimental and predicted lactate values of the consortium in the bioreactor.

TABLE 2 | Interaction parameters (ef$_{ji}$ in equation 8, Supplementary Material) found in co-cultures by the model.

	B. infantis	***B. vulgatus***	***E. coli***	***L. acidophilus***
B. infantis	–	99.99	16.33	45.05
B. vulgatus	−40.52	–	0.16	1.99
E. coli	−37.56	−57.14	–	−26.61
L. acidophilus	−70.23	−32.74	−99.99	–

Negative values indicate growth stimulation, and positive indicates a negative effect on growth. Values near 0 indicate a stronger effect.

the amounts of acetate and lactate appeared overestimated by the model (**Figures 4B**, **5B**).

As expected, growth of the consortium in the bioreactor resolved in a shorter time compared to the assays above (**Figure 2D**). Therefore, time was linearly adjusted for comparison and integration in the model. As in microplates, we observed a predominance of Bi and La. This observation was sustained during the course of the fermentation. Interestingly, the model was also able to predict this predominance (**Figure 2D**). In addition, both the model and data showed a full consumption of FOS at 12 h (**Figure 3D**). Finally, a good agreement of acetate and lactate amounts between the experimental evidence and the model was obtained (**Figures 4C**, **5C**). Since La was a good competitor during growth on FOS in the bioreactor, lactate concentrations appeared higher compared to previous experiments (**Figures 5A–C**). The parameters that define the production of lactate and acetate in Bi appear to be important in the four-bacterium co-culture, considering the predominance of Bi.

Finally, we performed a simple additional simulation to test the prediction capabilities of the model where a bacteriostatic agent is used against each member of the consortium (**Figure 6**). In every co-culture where Bi was able to grow, it predominated over the others (**Figures 6B–D**). On the other hand, if Bi was inhibited, Ec predominated in the co-culture (**Figure 6A**).

Parameter Identifiability Analysis

Parameter identifiability was used to find correlations between parameters (Parameter identifiability in Supplementary Material). This analysis is important for further reducing the number of fitted parameters by setting one of them and defining the other as a function. Inspection of the parameter covariance matrix is one way to find which parameters allow the model to be identifiable. As shown in **Figure 7**, highlighted cells display a high correlation (positive or negative). Usually parameters inside a cluster have a high correlation. In this case, this could be observed for all parameters from the same microorganism. For example, production of acetate and lactate in Bi are directly correlated, while some correlations between microorganisms were found. La's parameters (Y_{sx} – biomass yield, μ_{max} – Maximum growth rate, I_a – Acetate inhibition constant, I_l – Lactate inhibition constant) are inversely correlated to Ec bacterial parameters such as growth and inhibition constants. This suggests that the higher the La growth, the lower the *E. coli* biomass yield and higher inhibition. Several parameters associated to Bv growth were mostly directly correlated to Ec growth, indicating a more neutral or cooperative interaction.

Parameter Sensitivity Analysis

This analysis allows the determination of the influence of every parameter in each differential equation of the model. As shown in **Figure 8**, the effects of the parameters initially set are important in every ODE, due to the fact that Bi appears as the dominant microorganism in the consortium (**Figures 2C,D**). Specifically, the second parameter of the model (Bi's μ_{max}) has the highest influence on every other microorganism and their metabolic equations. Parameters K_3 and K_4 (Bi's inhibition constants of

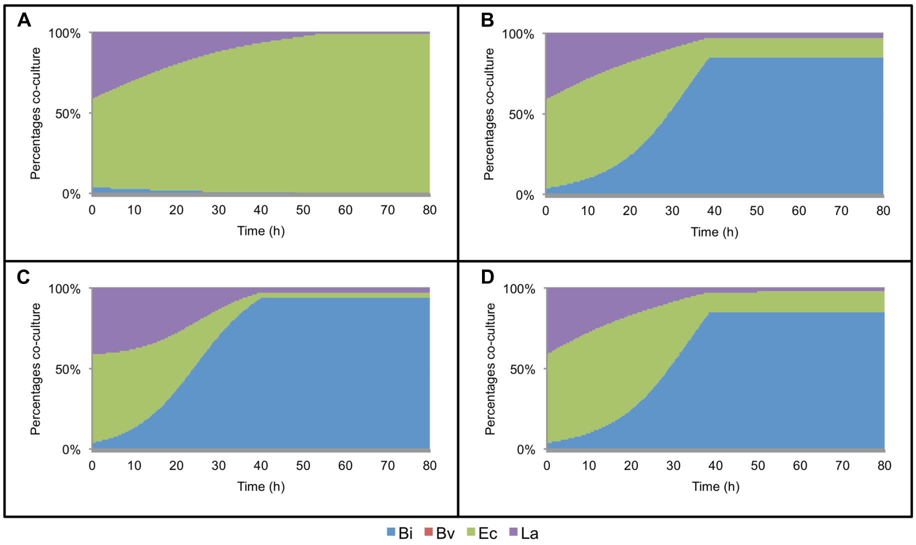


FIGURE 6 | Simulation of the effect of a bacteriostatic agents on the consortium. These agents are simulated to be directed and inhibit the growth of each member of the consortium. **(A)** *Bifidobacterium infantis* is unable to grow; **(B)** *Bacteroides vulgatus* is unable to grow; **(C)** *Escherichia coli* is unable to grow; **(D)** *Lactobacillus acidophilus* is unable to grow.

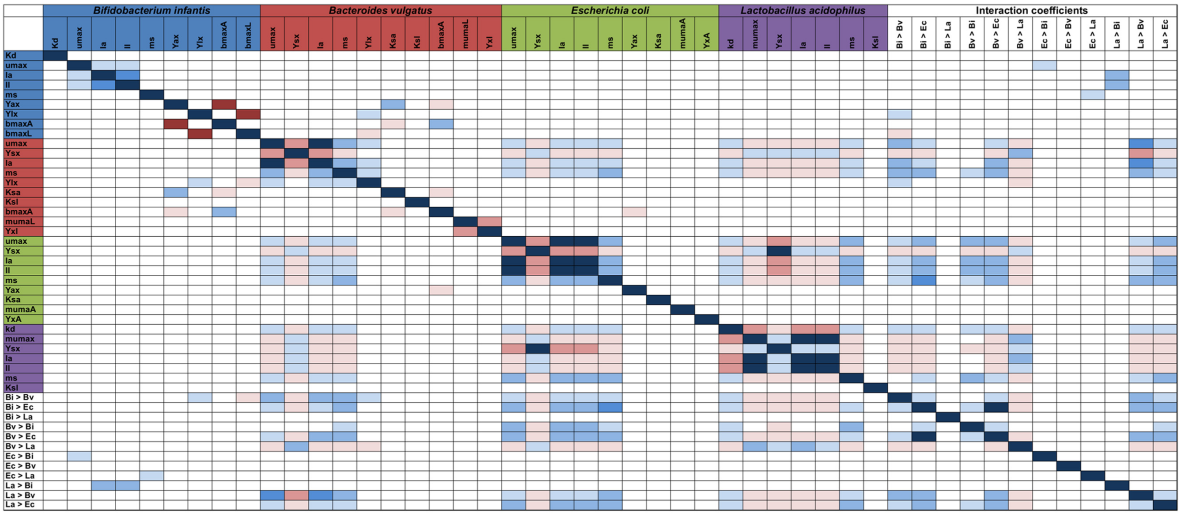


FIGURE 7 | Parameter model identifiability. Correlation values between each parameter in the model was calculated (for each microorganism including interaction). Only >|0.95| values are highlighted; red values are inversely correlated, while blue values are directly correlated. Parameters on both axes are indicated in **Table 1**.

acetate and lactate) also display a large influence on other microorganisms. In order to analyze the effects of the sensitive parameters found in the previous assay, **Figure 9** shows the average and standard deviation after 5000 iterations of randomly changing a parameter by 5% in its amount. The strongest effect of changing the value of Bi's $\mu_{\max}$ is on Ec cell numbers (**Figure 9A**), variable that can vary around 4% the value. On the other hand, a change in a parameter could also imply an advance or delay in the kinetics. **Figure 9B** shows the change in the FOS consumption kinetics due to effects of higher or lower values of Bi's lactate inhibition constant. Here we observed that changing the parameter only altered the dynamics of the ODE. Finally, **Figure 9C** shows the last case found in the sensitivity analysis, a parameter that is not sensitive to any differential equation. For example, measured Bv was not affected even after changing 50% parameter 25 (Ec substrate yield Y_{sx}).

DISCUSSION

The gut microbiome is a complex microbial community that modulates several host responses. This connection to host health makes it important to understand what forces guide microbiome composition and cause it to drift to an altered or dysbiotic

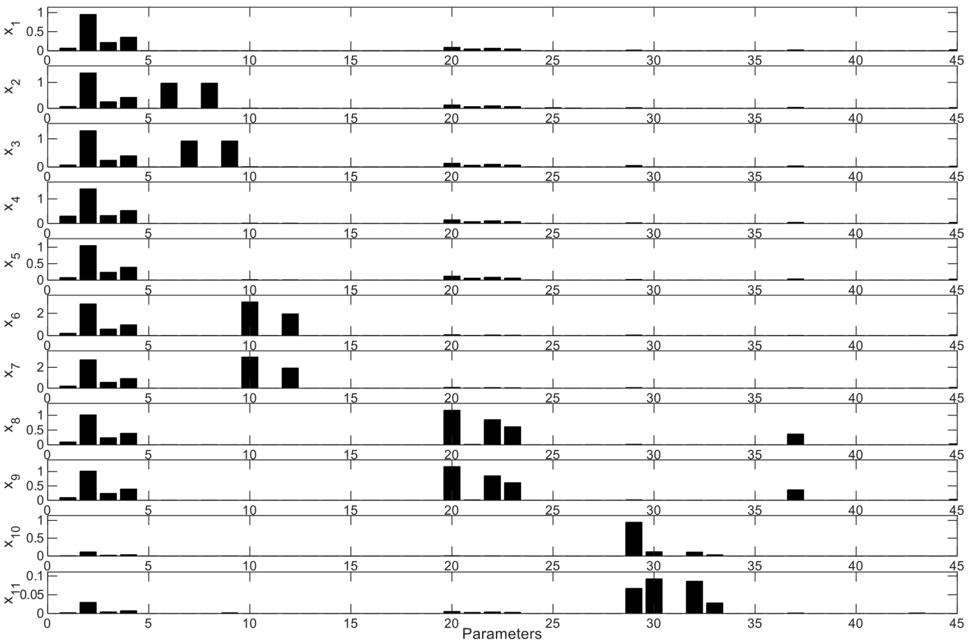


FIGURE 8 | Model parameter average sensitivity. Sensitivity (*Y*-axis) of each parameter (*X*-axis) for every ODE is shown. x_i represents every ODE described in the model ($x_1 = dS/dt$, $x_2 = dA/dt$, $x_3 = dL/dt$, $x_4 = dX_1/dt$, $x_5 = dX_{1m}/dt$, $x_6 = dX_2/dt$, $x_7 = dX_{2m}/dt$, $x_8 = dX_3/dt$, $x_9 = dX_{3m}/dt$, $x_{10} = dX_4/dt$, $x_{11} = dX_{4m}/dt$), where S: substrate; A: acetate; L: lactate. X: live biomass; X_m: total biomass. Parameters are in the same order in **Figure 6**.

microbiome (Cox et al., 2017). The interest in determining and predicting key factors in the establishment and maintenance of the gut microbiome is the major goal of several works (Trosvik et al., 2010a; Greenblum et al., 2012; Kettle et al., 2015; Shashkova et al., 2016).

Diet is a major modulator of the composition of the gut microbiome, and the nature of these substrates probably dictates which species predominate. In this study we evaluated if a mathematical model capturing metabolic interactions is able to recapitulate the composition and functions of a consortium of species of the gut microbiome. For this, we chose four representative bacteria of the infant gut microbiome, and using experimental data from mono and co-culture, a model was developed, calibrated and validated. Using a bioreactor, the developed model was assessed in a more controlled environment.

The system was studied during growth on FOS, a major prebiotic present in infant formula (Roberfroid et al., 2010). All members of the consortium display the ability to use this substrate (Roberfroid et al., 2010), including *E. coli* which could use small amounts of mono or disaccharides found in FOS. Moreover, different molecular mechanisms for FOS consumption have been described (Barrangou et al., 2003). In general the predictions by the model followed the *in vitro* behavior of the consortium, either in paired co-cultures, and growing the four-species consortium either in microplates or in a more controlled environment such as a biological reactor. This indicates that the model is able to predict changes in the bacterial abundance using only co-culture data for calibration.

It is very possible that interactions and parameters determined in this study are dependent on which prebiotic is used. FOS are commonly added to infant formula, but in combination with galactooligosaccharides (GOS), another important prebiotic (Garrido et al., 2013). Breast milk contains large concentrations of HMO, which are also a large catalog of oligosaccharides derived from lactose (Thomson et al., 2017). Moreover, the gut epithelium is covered with a mucin layer, containing oligosaccharides that could be used as carbon source by infant gut bacteria (Tailford et al., 2015). In a more realistic situation probably all these carbohydrates contribute to shape microbial interactions in different ways, since their chemical structure selects for specific microbial strains endowed with the cognate molecular machinery for utilization. However, if metabolic interactions are key in shaping microbiome composition, we could hypothesize that a mathematical model including these interactions could predict microbiome composition when other substrates are used.

We observed a good fit between experimental data and modeling results. This suggests that inhibitions observed in certain cases could be due to acetate and lactate production, variables that were quantified and included in the mathematical model. Both the reactor and the microplates had an initial pH of 5.5, however, pH was not regulated in the latter system. Considering this, similar results in both systems could also indicate that results obtained are independent of the pH.

A general good agreement was also observed for acid production and carbohydrate consumption. For Bi in mono-cultures and co-cultures where it predominates, the amounts of acetate and lactate produced are near a 3:2 ratio (Garrido et al., 2013). This was also observed during the growth of the consortium in the bioreactor. Acetate production by Ec was overestimated by the model (0.21 g of acetate per 1 g of FOS

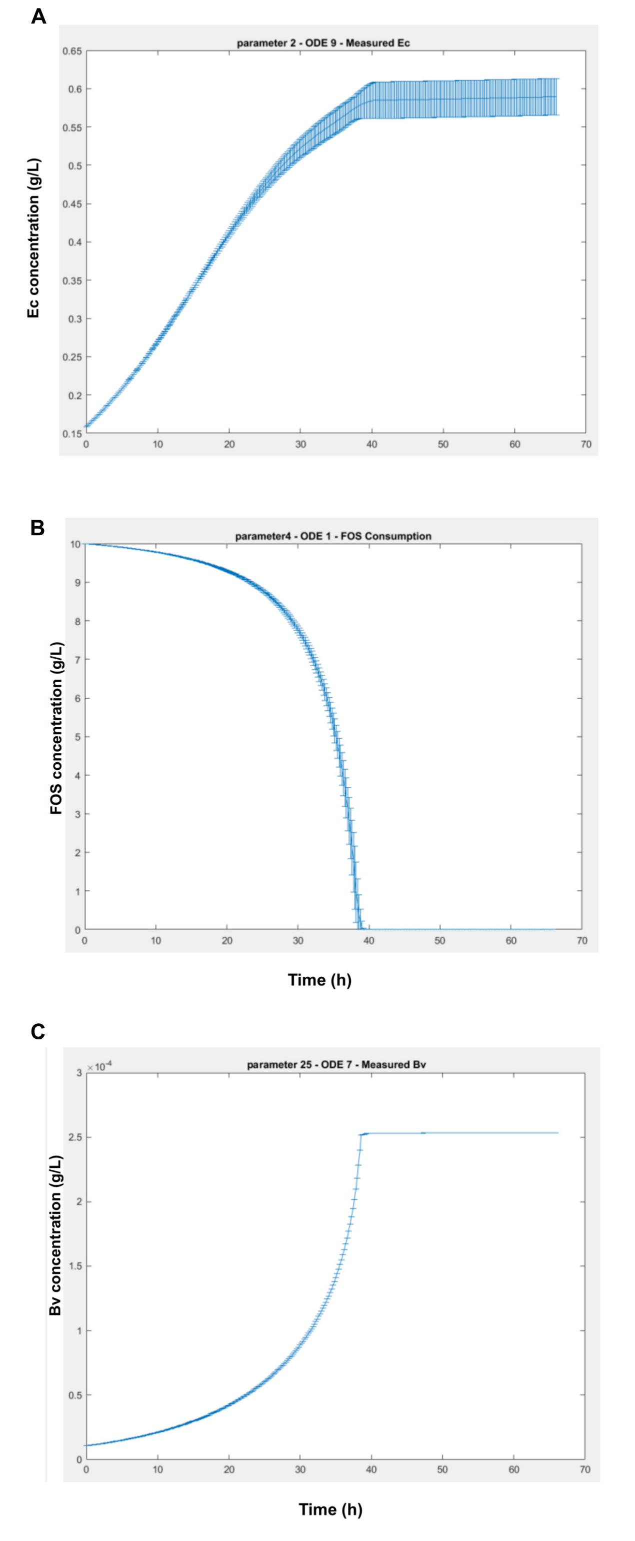


FIGURE 9 | Variation of the ODEs values (g/L) over time due a 5% change in the parameters in 5000 iterations. **(A)** worst case scenario, with parameter 2 (Bi's μ_{max}) affecting measured Ec ODE; **(B)** a change in kinetics scenario, with parameter 4 (Bi's lactate inhibition constant), affecting FOS consumption; **(C)** a non-sensitive parameter (measured Bv ODE), for example to Ec substrate yield.

consumed). In general Ec was thought to benefit from other microorganism activities in that it uses mono or disaccharides released to the media (**Table 2**) (Ravcheev et al., 2013; Vuoristo et al., 2015). Another possibility might be protein fermentation by Ec (Lulit and Strohl, 1990). Lactate production of La determined by the model was around 0.63 g per 1 g of substrate, a similar yield in lactose reported (Fu and Mathews, 1999).

In some co-cultures, the concentration of either acetate or lactate was overestimated. This was evident in Bi co-cultures whenever it predominated. Parameters of the model could be much better estimated in experiments with improved resolution and more frequent measurements. Since the model in co-cultures defines the intervals where the parameters are most sensitive, it is possible that an increase in the number of samples would reduce the variation of underestimated parameters. The time points where the substrate is being fully consumed are critical, and microorganisms could find another substrate for growth (for secondary fermenters) or entering to a stationary phase. Also, for *Bacteroides* and *Escherichia* cultures, the microbial concentration could be overestimated by some intrinsic pathways of these genera (Neis et al., 2015; Vuoristo et al., 2015).

Moreover, while acetate and lactate are major metabolic products in this system, a more complete picture could be obtained if the model included other metabolites. Adding more equations of utilization and inhibition by metabolites such as ethanol, propionate, butyrate and amino acids could be important. Amino acid cross-feeding between *Bacteroides* and *Lactobacillus* supports bacterial growth *in vitro* and *in silico* (Magnúsdóttir et al., 2016).

The analysis of bacteriostatic agent effects on the culture suggested that Bi should be predominant if other bacteria are inhibited. However, when Bi is inhibited, La or Bv should grow more than Ec, because of their glycolytic properties (Ravcheev et al., 2013). This is a limitation of the model, probably due to missing functions that describe the breakdown of complex carbohydrates by Bv, or the protein fermentation as a carbon source of bacteria. In addition, further work could corroborate these hypotheses by adding the respective antibiotic and measuring the same variables used in this work.

A possible application of this initial ODE-based model is that it could be used to predict microbial composition in the gut based on diet, at least in simpler microbiome communities. This work indicates that it is possible to have a good approach to this goal if metabolic interactions are included. Moreover, bacterial composition of a microbiome could eventually be optimized, for example to increase production of acetate and lactate. These two acids are important modulators of health outcomes in the gut. For example acetate has been shown to prevent pathogen colonization (Fukuda et al., 2011), and lactate in the adult gut microbiome is used by butyrate-producing bacteria (Moens et al., 2016), a health-promoting SCFA (Louis and Flint, 2017).

Finally, this model could be useful to study interactions using a more complex set of species of gut microbiome species. In general these results could be important to predict the composition of microbial communities where metabolic interactions are relevant. Considering the flexibility of incorporating product equations and growth inhibitions to the model, this model could be used to find microbial consortia with desired metabolic properties such as maximized acid production.

AUTHOR CONTRIBUTIONS

FP and DM performed experiments. FP developed the model. DG provided materials and reagents. JP-C and DG conceived the experiments, wrote, and edited the manuscript.

FUNDING

This work was funded by Fondecyt de Iniciacion 11130518, Fondecyt de Postdoctorado 3160525 and FONDEF IDEA ID16i10045.

ACKNOWLEDGMENTS

We thank Dr. Eduardo Agosin for the support with equipment and methods, Paulina Torres for supporting in bioreactor cultures. We thank Dr. Martin Gotteland and Dr. Patricia García Cañete for providing materials for this study. We also thank Dr. David Mills and Lucy Joseph for access to V&E Culture Collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02507/full#supplementary-material

REFERENCES

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703. doi: 10.1016/j.chom.2015.04.004

Barrangou, R., Altermann, E., Hutkins, R., Cano, R., and Klaenhammer, T. R. (2003). Functional and comparative genomic analyses of an operon involved in fructooligosaccharide utilization by *Lactobacillus acidophilus*. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8957–8962. doi: 10.1073/pnas.1332765100

Bartosch, S., Fite, A., Macfarlane, G. T., and McMurdo, M. E. T. (2004). Characterization of bacterial communities in feces from healthy elderly volunteers and hospitalized elderly patients by using real-time PCR and effects of antibiotic treatment on the fecal microbiota. *Appl. Environ. Microbiol.* 70, 3575–3581. doi: 10.1128/AEM.70.6.3575-3581.2004

Cameron, E. A., Kwiatkowski, K. J., Lee, B.-H., Hamaker, B. R., Koropatkin, N. M., and Martens, E. C. (2014). Multifunctional nutrient-binding proteins adapt human symbiotic bacteria for glycan competition in the gut by separately promoting enhanced sensing and catalysis. *mBio* 5:e01441-14. doi: 10.1128/mBio.01441-14

Cinquin, C., Le Blay, G., Fliss, I., and Lacroix, C. (2004). Immobilization of infant fecal microbiota and utilization in an in vitro colonic fermentation model. *Microb. Ecol.* 48, 128–138. doi: 10.1007/s00248-003-2022-7

Cox, L. M., Yamanishi, S., Sohn, J., Alekseyenko, A. V., Leung, J. M., Cho, I., et al. (2017). Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* 158, 705–721. doi: 10.1016/j.cell.2014.05.052

Egan, M., Motherway, M. O. C., Ventura, M., and van Sinderen, D. (2014a). Metabolism of sialic acid by *Bifidobacterium breve* UCC2003. *Appl. Environ. Microbiol.* 80, 4414–4426. doi: 10.1128/AEM.01114-14

Egan, M., O'Connell Motherway, M., Kilcoyne, M., Kane, M., Joshi, L., Ventura, M., et al. (2014b). Cross-feeding by *Bifidobacterium breve* UCC2003 during co-cultivation with *Bifidobacterium bifidum* PRL2010 in a mucin-based medium. *BMC Microbiol.* 14:282. doi: 10.1186/s12866-014-0282-7

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

Fu, W., and Mathews, A. P. (1999). Lactic acid production from lactose by *Lactobacillus plantarum*: kinetic model and effects of pH, substrate, and oxygen. *Biochem. Eng. J.* 3, 163–170. doi: 10.1016/S1369-703X(99)00014-5

Fukuda, S., Toh, H., Hase, K., Oshima, K., Nakanishi, Y., Yoshimura, K., et al. (2011). Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* 469, 543–547. doi: 10.1038/nature09646

Garrido, D., Ruiz-moyano, S., Jimenez-espinoza, R., and Eom, H. (2013). Utilization of galactooligosaccharides by *Bifidobacterium longum* subsp. *infantis* isolates. *Food Microbiol.* 33, 262–270. doi: 10.1016/j.fm.2012.10.003

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

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U.S.A.* 109, 594–599. doi: 10.1073/pnas.1116053109

Hastie, T., and Tibshirani, R. (1990). Generalized additive models. *Stat. Sci.* 10, 354–363. doi: 10.2307/2246134

Hecht, A. L., Casterline, B. W., Earley, Z. M., Goo, Y. A., Goodlett, D. R., Bubeck Wardenburg, J., et al. (2016). Strain competition restricts colonization of an enteric pathogen and prevents colitis. *EMBO Rep.* 94:e201642282. doi: 10.15252/embr.201642282

Kanehisa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–D357. doi: 10.1093/nar/gkj102

Kassinen, A., Malinen, E., Krogius, L., Palva, A., and Rinttila, T. (2004). Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. *J. Appl. Microbiol.* 97, 1166–1177. doi: 10.1111/j.1365-2672.2004.02409.x

Kettle, H., Louis, P., Holtrop, G., Duncan, S. H., and Flint, H. J. (2015). Modelling the emergent dynamics and major metabolites of the human colonic microbiota. *Environ. Microbiol.* 17, 1615–1630. doi: 10.1111/1462-2920.12599

Louis, P., and Flint, H. J. (2017). Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* 19, 29–41. doi: 10.1111/1462-2920.13589

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220–230. doi: 10.1038/nature11550

Lulit, G. W., and Strohl, W. R. (1990). Comparison of growth, acetate production, and acetate inhibition of *Escherichia coli* strains in batch and fed-batch fermentations. *Appl. Environ. Microbiol.* 56, 1004–1011.

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2016). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35, 81–89. doi: 10.1038/nbt.3703

Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D. A., Hirschfield, G. M., Hold, G., et al. (2016). The gut microbiota and host health: a new clinical frontier. *Gut* 65, 330–339. doi: 10.1136/gutjnl-2015-309990

Medina, D., Pinto, F., Ovalle, A., Thomson, P., and Garrido, D. (2017). Prebiotics mediate microbial interactions in a consortium of the infant gut microbiome. *Int. J. Mol. Sci.* 18:E2095. doi: 10.3390/ijms18102095

Mendoza, S. N., Cañón, P. M., Contreras, A., Ribbeck, M., and Agosin, E. (2017). Genome-scale reconstruction of the metabolic network in *Oenococcus oeni* to assess wine malolactic fermentation. *Front. Microbiol.* 8:534. doi: 10.3389/FMICB.2017.00534

Moens, F., Weckx, S., and De Vuyst, L. (2016). Bifidobacterial inulin-type fructan degradation capacity determines cross-feeding interactions between bifidobacteria and Faecalibacterium prausnitzii. *Int. J. Food Microbiol.* 231, 76–85. doi: 10.1016/j.ijfoodmicro.2016.05.015

Neis, E. P. J. G., Dejong, C. H. C., and Rensen, S. S. (2015). The role of microbial amino acid metabolism in host metabolism. *Nutrients* 7, 2930–2946. doi: 10.3390/nu7042930

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, S., Manichanh, C., et al. (2013). A human gut microbial gene catalog established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821.A

Rajilić-Stojanović, M., and de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* 38, 996–1047. doi: 10.1111/1574-6976.12075

Rakoff-Nahoum, S., Foster, K. R., and Comstock, L. E. (2016). The evolution of cooperation within the gut microbiota. *Nature* 533, 255–259. doi: 10.1038/nature17626

Ravcheev, D. A., Godzik, A., Osterman, A. L., and Rodionov, D. A. (2013). Polysaccharides utilization in human gut bacterium *Bacteroides thetaiotaomicron*: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics* 14:873. doi: 10.1186/1471-2164-14-873

Roberfroid, M., Gibson, G. R., Hoyles, L., McCartney, A. L., Rastall, R., Rowland, I., et al. (2010). Prebiotic effects: metabolic and health benefits. *Br. J. Nutr.* 104, S1–S63. doi: 10.1017/S0007114510003363

Rogowski, A., Briggs, J. A., Mortimer, J. C., Tryfona, T., Terrapon, N., Lowe, E. C., et al. (2015). Glycan complexity dictates microbial resource allocation in the large intestine. *Nat. Commun.* 6:7481. doi: 10.1038/ncomms8481

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

Sacher, J., Saa, P., Cárcamo, M., López, J., Gelmi, C. A., and Pérez-Correa, R. (2011). Improved calibration of a solid substrate fermentation model. *Electron. J. Biotechnol.* 14, 7. doi: 10.2225/vol14-issue5-fulltext-7

Sevelsted, A., Stokholm, J., Bonnelykke, K., and Bisgaard, H. (2015). Cesarean section and chronic immune disorders. *Pediatrics* 135, e92–e98. doi: 10.1542/peds.2014-0596

Shashkova, T., Popenko, A., Tyakht, A., Peskov, K., Kosinsky, Y., Bogolubsky, L., et al. (2016). Agent based modeling of human gut microbiome interactions and perturbations. *PLOS ONE* 11:e0148386. doi: 10.1371/journal.pone.0148386

Tailford, L. E., Crost, E. H., Kavanaugh, D., and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. *Front. Genet.* 6:81. doi: 10.3389/fgene.2015.00081

Tamburini, S., Shen, N., Wu, H. C., and Clemente, J. C. (2016). The microbiome in early life: implications for health outcomes. *Nat. Med.* 22, 713–722. doi: 10.1038/nm.4142

Thiele, I., Heinken, A., and Fleming, R. M. T. (2013). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

Thiele, I., and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi: 10.1038/nprot.2009.203.A

Thiele, I., Vlassis, N., and Fleming, R. M. T. (2014). FASTGAPFILL: efficient gap filling in metabolic networks. *Bioinformatics* 30, 2529–2531. doi: 10.1093/bioinformatics/btu321

Thomson, P., Medina, D. A., and Garrido, D. (2017). Human milk oligosaccharides and infant gut bifidobacteria: molecular strategies for their utilization. *Food Microbiol.* (in press). doi: 10.1016/j.fm.2017.09.001

Trosvik, P., Rudi, K., Strætkvern, K. O., Jakobsen, K. S., Næs, T., and Stenseth, N. C. (2010a). Web of ecological interactions in an experimental gut microbiota. *Environ. Microbiol.* 12, 2677–2687. doi: 10.1111/j.1462-2920.2010.02236.x

Trosvik, P., Stenseth, N. C., and Rudi, K. (2010b). Convergent temporal dynamics of the human infant gut microbiota. *ISME J.* 4, 151–158. doi: 10.1038/ismej.2009.96

Tuncil, Y. E., Xiao, Y., Porter, N. T., Reuhs, B. L., Martens, E. C., and Hamaker, B. R. (2017). Reciprocal prioritization to dietary glycans by gut bacteria in a competitive environment promotes stable coexistence. *mBio* 8:e01068-17. doi: 10.1128/mBio.01068-17

Tuomivaara, S. T., Yaoi, K., O'Neill, M. A., and York, W. S. (2015). Generation and structural validation of a library of diverse xyloglucan-derived oligosaccharides, including an update on xyloglucan nomenclature. *Carbohydr. Res.* 402, 56–66. doi: 10.1016/j.carres.2014.06.031

Vogt, S. L., Peña-Díaz, J., and Finlay, B. B. (2015). Chemical communication in the gut: effects of microbiota-generated metabolites on gastrointestinal bacterial pathogens. *Anaerobe* 34, 106–115. doi: 10.1016/j.anaerobe.2015.05.002

Vuoristo, K. S., Mars, A. E., Sangra, J. V., Springer, J., Eggink, G., and Sanders, J. P. M. (2015). Metabolic engineering of the mixed-acid fermentation pathway of *Escherichia coli* for anaerobic production of glutamate and itaconate. *AMB Express* 5:61. doi: 10.1186/s13568-015-0147-y

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* 5, 220–230. doi: 10.1038/ismej.2010.118

Ward, T. (2014). The Information Theoretically Efficient Model (ITEM): a model for computerized analysis of large datasets. arXiv:1409.6075

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70, 495–518. doi: 10.1111/j.1467-9868.2007.00646.x

Zhang, G., Mills, D. A., and Block, D. E. (2009). Development of chemically defined media supporting high-cell-density growth of lactococci, enterococci, and streptococci. *Appl. Environ. Microbiol.* 75, 1080–1087. doi: 10.1128/AEM.01416-08

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 Microbiology

METHODS
published: 28 March 2018
doi: 10.3389/fmicb.2018.00509

Variance Component Selection With Applications to Microbiome Taxonomic Data

Jing Zhai[1], *Juhyun Kim*[2], *Kenneth S. Knox*[3], *Homer L. Twigg III*[4], *Hua Zhou*[2] and *Jin J. Zhou*[1]*

[1] Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States, [2] Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, United States, [3] Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Department of Medicine, University of Arizona, Tucson, AZ, United States, [4] Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University Medical Center, Indianapolis, IN, United States

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Gwenael Piganeau,
FR3724 Observatoire Océanologique
de Banyuls sur Mer (OOB), France
Jun Chen,
Mayo Clinic, United States

***Correspondence:**
Jin J. Zhou
jzhou@email.arizona.edu

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

Received: 15 November 2017
Accepted: 06 March 2018
Published: 28 March 2018

Citation:
Zhai J, Kim J, Knox KS, Twigg HL III,
Zhou H and Zhou JJ (2018) Variance
Component Selection With
Applications to Microbiome
Taxonomic Data.
Front. Microbiol. 9:509.
doi: 10.3389/fmicb.2018.00509

High-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. Microbiome data are summarized as counts or composition of the bacterial taxa at different taxonomic levels. An important problem is to identify the bacterial taxa that are associated with a response. One method is to test the association of specific taxon with phenotypes in a linear mixed effect model, which incorporates phylogenetic information among bacterial communities. Another type of approaches consider all taxa in a joint model and achieves selection via penalization method, which ignores phylogenetic information. In this paper, we consider regression analysis by treating bacterial taxa at different level as multiple random effects. For each taxon, a kernel matrix is calculated based on distance measures in the phylogenetic tree and acts as one variance component in the joint model. Then taxonomic selection is achieved by the lasso (least absolute shrinkage and selection operator) penalty on variance components. Our method integrates biological information into the variable selection problem and greatly improves selection accuracies. Simulation studies demonstrate the superiority of our methods versus existing methods, for example, group-lasso. Finally, we apply our method to a longitudinal microbiome study of Human Immunodeficiency Virus (HIV) infected patients. We implement our method using the high performance computing language `Julia`. Software and detailed documentation are freely available at https://github.com/JingZhai63/VCselection.

Keywords: Human Immunodeficiency Virus (HIV), lasso, longitudinal study, lung microbiome, MM-algorithm, variance component models, variable selection

1. INTRODUCTION

The advent of high-throughput sequencing technologies has produced extensive microbial community data, which reveals the impact of human microbes on health and various diseases (Mardis, 2008; Haas et al., 2011; Hodkinson and Grice, 2015; Kuleshov et al., 2016; Wang and Jia, 2016). Microbial community data collected from oral, skin, and gastrointestinal tract samples have received early attention (Eckburg et al., 2005; Gill et al., 2006; Turnbaugh et al., 2009; Dewhirst et al., 2010; Grice and Segre, 2011). Studies of the respiratory tract microbiome did not start until the discovery of microbiome in the lungs of both healthy (Erb-Downward et al., 2011; Morris et al., 2013; Twigg III et al., 2013) and diseased populations (Zemanick et al., 2011; Lozupone et al., 2013) using culture-independent techniques. A pulmonary microbiome dataset was sampled longitudinally from 30 HIV-infected individuals after starting highly active antiretroviral therapy (HAART). The objective is to study how the pulmonary microbiome impacts lung function of advanced HIV patients after HAART (Garcia et al., 2013; Lozupone et al., 2013; Twigg III et al., 2016).

After microbiome sequences have been acquired, they are usually clustered into Operational Taxonomic Units (OTUs): groups of sequences that correspond to taxonomic clusters or monophyletic groups (Caporaso et al., 2010). The abundance of an OTU is defined as the number of sequences in that OTU. The microbial community is then described by a list of OTUs, their abundances, and a phylogenetic tree. Regression methods have been a powerful tool to identify clusters of OTUs that are associated with or predictive of host phenotypes (Zhao et al., 2015; Wang and Zhao, 2016; Wang et al., 2017). Microbiome data presents several challenges. First microbiome abundances are sparse and the number of OTUs is usually much bigger than sample size. In our longitudinal data set, there are 2,964 OTUs and only two of them have abundance greater than 5%. When OTUs are included as predictors for clinical phenotypes in a regression model, regularizations are often used to overcome ill-conditioning. For example, Lin et al. (2014) proposed a linear log-contrast model with ℓ_1 regularization. Another possible strategy to overcome the sparsity of microbial data is to cluster multiple OTUs into their higher phylogenetic levels, e.g., genus, order, and phylum. Shi et al. (2016) extended Lin et al.'s (2014) model to allow selecting taxa at different higher taxonomic ranks. However, both methods overlook the distance information in the phylogenetic tree. A network-constrained sparse regression is proposed to achieve better prediction performance through a Laplacian regularization (Chen et al., 2012b, 2015b). Another popular approach for sparse linear regression is the group-wise selection scheme, group-lasso, which selects an entire group for inclusion or exclusion (Yuan and Lin, 2006; Garcia et al., 2013; Simon et al., 2013; Yang and Zou, 2015). Therefore, group-lasso is a natural tool for incorporating group information defined by the phylogenetic tree, but still misses fine level information. To encourage hierarchically close species to have similar effects on the phenotype, Wang and Zhao (2016) and Wang et al. (2017) both used tree topology information and fused variables that stay closer in a tree. However, this assumption may be violated. For example, the bacteria *Clostridia*, some species in this class convert dietary fiber into anti-inflammatory short-chain fatty acids, while others cause severe colitis. We, therefore, need a method that can incorporate biologically meaningful cluster information, phylogenetic distance, or tree information, can encourage sparse feature selection, and can handle possible adverse effect within clusters.

By modeling microbiome cluster effects as random effects, Zhai et al. (2017b) proposed a variance component model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \sum_{l}^{L} h_l + \boldsymbol{\varepsilon}$$
$$\boldsymbol{b} \sim \mathcal{N}(0, \sigma_d^2 \boldsymbol{I}_n),\ h_l \sim \mathcal{N}(\boldsymbol{0}, \sigma_{gl}^2 \boldsymbol{K}_l),\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_n), \quad (1)$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\varepsilon}$ are the vertically stacked vectors/matrices of $\boldsymbol{y}_i$, $\boldsymbol{X}_i$, and $\boldsymbol{\varepsilon}_i$. The $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of n_i repeated measures of the quantitative phenotype for an individual i. $\boldsymbol{X}_i$ is the $n_i \times p$ covariates. The $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of the random error. $\boldsymbol{Z}_i = (1, \ldots, 1)'$ is an $n_i \times 1$ design matrix linking the vector of random effects b_i to $\boldsymbol{y}_i$. $\boldsymbol{Z}$ is a block diagonal matrix with $\boldsymbol{Z}_i$ on its diagonal. $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects. The $\boldsymbol{b} = (b_i)$ is the subject-specific random effects. L is the total number of microbiome taxonomic clusters, N is the total number of individuals and $\sum_{i=1}^{N} n_i$ is the total number of observations. In model (Equation 1), h_l is the random effects generated by microbiome taxa l with covariance $\sigma_{gl}^2 \boldsymbol{K}_l$. $\boldsymbol{K}_l$ is a positive-definite kernel matrix derived from a distance matrix that is calculated based on the OTU abundances of taxa in the phylogenetic tree. Two common distance matrices are UniFrac Distance (Lozupone and Knight, 2005) and Bray-Curtis dissimilarity (Bray and Curtis, 1957). Therefore,

$$\mathrm{Var}(\boldsymbol{y}) = \sigma_d^2 \boldsymbol{Z}'\boldsymbol{Z} + \sum_{l=1}^{L} \sigma_{gl}^2 \boldsymbol{K}_l + \sigma_e^2 \boldsymbol{I}_n, \quad (2)$$

where σ_{gl}^2 and σ_d^2 are the phenotypic variance from microbiome clusters and between subject variance from repeated measurements. σ_e^2 is the within-subject variance that cannot be explained by either microbiome or repeated measurements. To identify associated microbiome taxa at different phylogeny levels is to select non-zero variance components at different phylogeny levels.

In this article, we adopt a penalized likelihood approach by regularizing variance components based on linear mixed effect models: variance component lasso selection (VC-lasso). We incorporate the phylogenetic tree information by using kernel matrices. We reduce the dimensionality of large and very sparse OTU abundances within a cluster by translating them into a random effect. Furthermore, our method can be applied to a longitudinal design, where an unpenalized variance component that captures the correlation of repeated measurements is included. Our Majorization-Minimization (MM) algorithm for variance component selection guarantees estimation and selection computational efficiency (Hunter and Lange, 2004; Hunter and Li, 2005; Zhou et al., 2011, 2015; Lange, 2016). Many statistical methods have been proposed related to the selection of random effects. Ibrahim et al. (2011) considered jointly selecting fixed and random effect in mixed effect model using the maximum likelihood with the smoothly clipped absolute deviation (SCAD) and adaptive lasso penalization. Fan and Li (2012) proposed a group variable selection strategy to select and estimate important random effects. Hui et al. (2017) extended this strategy to generalized linear mixed model by combining the penalized quasi-likelihood (PQL) estimation with sparsity-inducing penalties on the fixed and random coefficients. However, none of these methods can be easily extended to microbiome data and none of them use variance component regularization.

The rest of this paper is organized as follows. We introduce the variance component lasso selection method in section 2.

Section 3 conducts comparative simulation studies. Section 4 presents simulation and real data analysis results. We conclude with a discussion in section 5.

2. METHODS

2.1. Lasso Penalized Log-Likelihood

We consider model (Equation 2) with model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_m^2)$. The log-likelihood of our model is:

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3}$$

where

$$\boldsymbol{V} = \sum_{i=1}^{m} \sigma_i^2 \boldsymbol{V}_i.$$

For the selection of non-zero variance components among a large number of variance components, we estimate the regression parameter $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ by minimizing the lasso penalized log-likelihood function

$$pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \boldsymbol{y}, \boldsymbol{X}, \lambda) = -L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^{m} c_i \sigma_i, \tag{4}$$

subject to nonnegativity constraint $\sigma_i \geq 0$. The first part $-L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ of the penalized function (Equation 4) is the negative log-likelihood defined in Equation (3). The second part is the lasso penalty to enforce shrinkage of high-dimensional components. We do not penalize fixed effects $\boldsymbol{\beta}$. λ is the tuning parameter controlling model complexity; $c_i \in \{0, 1\}$ allows differential shrinkage of specific variance components. For example, when modeling longitudinal phenotypes with random intercept model, the corresponding variance component is unpenalized and always stays in the model. c_i can be chosen using different weighting schemes based on prior knowledge such as functional annotations.

2.2. Minimization of Penalized Likelihood via MM Algorithm

Minimizing the penalized negative log-likelihood is challenging due to non-convexity. Based on the Majorization-Minimization (MM) algorithm (Lange et al., 2000; Hunter and Lange, 2004), Zhou et al. (2015) proposed a strategy for maximizing the log-likelihood Equation (3) by alternate updating $\boldsymbol{\beta}$ and variance components $\boldsymbol{\sigma}^2$. We follow the same strategy to solve the lasso penalized likelihood estimation problem (Algorithm 1).

Algorithm 1: MM algorithm for minimizing lasso penalized likelihood (Equation 4).

```
Data: $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{V}_1, \dots, \boldsymbol{V}_m, \lambda$
Result: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2$ such that $pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^{m} c_i \sigma_i$ is minimized.
Initialize $\sigma_i^{(0)} > 0.\ i = 1, \dots, m$ repeat
    $\boldsymbol{V}^{(t)} \leftarrow \sum_{i=1}^{m} \sigma_i^{2(t)} \boldsymbol{V}_i$;
    $\boldsymbol{\beta}^{(t)} \leftarrow \mathrm{argmin}_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$;
    $\sigma_i^{(t+1)} \leftarrow \sigma_i^{(t)}$ by finding polynomial roots of $\boldsymbol{P}(\cdot) = 0,\ i = 1, \dots, m$
    $\boldsymbol{P}(\sigma_i^{(t+1)}) = \sigma_i^{4(t+1)} \mathrm{tr}(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i) + \lambda \sigma_i^{3(t+1)} - \sigma_i^{4(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})'\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})$
until objective function pl converges;
```

Given $\boldsymbol{\sigma}^{2(t)}$, updating $\boldsymbol{\beta}$ is a general least squares problem with solution

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}'\boldsymbol{V}^{-(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-(t)}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{V}^{-(t)}$ represents the tth-step update of $\boldsymbol{V}^{-1}$. Given $\boldsymbol{\beta}^{(t)}$, updating the variance components $\boldsymbol{\sigma}^2$ invokes the MM principle. To minimize the objective function $pl(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$, the majorization step operates by creating a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ that satisfies two conditions

$$\begin{aligned} \text{dominance condition}: pl(\boldsymbol{\theta}) &\leq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \text{ for all } \boldsymbol{\theta} \\ \text{tangent condition}: pl(\boldsymbol{\theta}^{(t)}) &= g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}). \end{aligned}$$

The second M of the MM principle minimizes the surrogate function to produce the next iterate $\boldsymbol{\theta}^{(t+1)}$. Then we have

$$pl(\boldsymbol{\theta}^{(t+1)}) \leq g(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \leq g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = pl(\boldsymbol{\theta}^{(t)}).$$

Therefore, when the surrogate function is minimized, the objective function $f(\boldsymbol{\theta})$ is driven downhill. We combine two following majorizations to construct the surrogate function. First, with all $\boldsymbol{V}_i$ being positive semidefinite, Zhou et al. (2015) show that

$$\begin{aligned} \boldsymbol{V}^{(t)}\boldsymbol{V}^{-1}\boldsymbol{V}^{(t)} &= \left(\sum_{i=1}^{m} \sigma_i^{2(t)}\boldsymbol{V}_i\right)\left(\sum_{i=1}^{m} \sigma_i^2\boldsymbol{V}_i\right)^{-1}\left(\sum_{i=1}^{m} \sigma_i^{2(t)}\boldsymbol{V}_i\right) \\ &\preceq \sum_{i=1}^{m} \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}}\left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}}\sigma_i^{2(t)}\boldsymbol{V}_i\right) \\ &\quad \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}}\sigma_i^2\boldsymbol{V}_i\right)^{-1}\left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}}\sigma_i^{2(t)}\boldsymbol{V}_i\right) \\ &= \sum_{i=1}^{m} \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\boldsymbol{V}_i^{-1}\boldsymbol{V}_i = \sum_{i=1}^{m} \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i, \end{aligned}$$

leading to the first majorization

$$\begin{aligned} &(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &\preceq (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-(t)}\left(\sum_{i=1}^{m} \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\right)\boldsymbol{V}^{-(t)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{6}$$

It separates the variance components $\sigma_1^2, \dots, \sigma_m^2$ in the quadratic term of the log-likelihood function (Equation 4). By the supporting hyperplane inequality, the second majorization is

$$\ln \det \boldsymbol{V} \leq \ln \det \boldsymbol{V}^{(t)} + \mathrm{tr}\big[\boldsymbol{V}^{-(t)}\big(\boldsymbol{V} - \boldsymbol{V}^{(t)}\big)\big], \tag{7}$$

which separates $\sigma_1^2, \dots, \sigma_m^2$ in the log-determinant term of Equation (4). The overall majorization $g(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)})$ of $pl(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ is obtained by combining Equations (6) and (7)

$$\begin{aligned} g\big(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)}\big) &= \frac{1}{2}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}\big) + \frac{1}{2}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)'\boldsymbol{V}^{-(t)} \\ &\quad \left(\sum_{i=1}^{m} \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{V}_i\right)\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) + \lambda\sum_{i=1}^{m}\sigma_i + s^{(t)} \\ &= \sum_{i=1}^{m}\Bigg[\frac{\sigma_i^2}{2}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\big) + \frac{\sigma_i^{4(t)}}{2\sigma_i^2}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)' \\ &\quad \boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) + \lambda\sigma_i\Bigg] + s^{(t)}, \end{aligned} \tag{8}$$

where $s^{(t)}$ is an irrelevant constant term.

We minimize the surrogate function (Equation 8) by setting the derivative of $g(\boldsymbol{\sigma}^2|\boldsymbol{\sigma}^{2(t)})$ to zero. The update $\sigma_i^{(t+1)}$ for variance component $\sigma_i^{(t)}$ is chosen among the positive roots of the polynomial

$$\begin{aligned} \boldsymbol{P}\left(\sigma_i^{(t+1)}\right) &= \sigma_i^{4(t+1)}\mathrm{tr}\big(\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\big) + \lambda\sigma_i^{3(t+1)} \\ &\quad - \sigma_i^{4(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big)'\boldsymbol{V}^{-(t)}\boldsymbol{V}_i\boldsymbol{V}^{-(t)}\big(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\big) \end{aligned}$$

or 0, whichever yields the largest objective value. The alternating updates repeat until

$$|\, pl\big(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\sigma}^{2(t+1)}\big) - pl\big(\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}\big)\,| < tol * (|\, pl\big(\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{2(t)}\big)\,| + 1),$$

where *tol* is the pre-specified tolerance. The default tolerance is 10^{-4}.

2.3. Tuning Parameter Selection

The tuning parameter λ in the penalized likelihood estimation is chosen by a 5-fold cross-validation procedure based on g-Measure $= \sqrt{\text{sensitivity} * \text{specificity}}$. g-Measure is an indicator of the model selection accuracy. g-Measure $= 1$ indicates the best accuracy and g-Measure $= 0$ the worst (Zhai et al., 2017a). It can counteract the imbalance between the number of of irrelevant and relevant clusters. Therefore, we present g-Measure instead of sensitivity (true positive rate) and specificity (true negative rate) alone (Supplementary Material section 3). Akaike Information Criterion (AIC) (Akaike, 1998) and Schwarzs Bayesian Information Criterion (BIC) (Schwarz et al., 1978) are used in the real data analysis. Performance comparisons between cross-validation and AIC/BIC are provided in the Supplementary Material section 4.

2.4. Software Implementation

We implement our method using the high performance computing language `Julia`. UniFrac distance matrices are computed using our `Julia` package `PhylogeneticDistance`.

TABLE 1 | Simulation parameter configurations.

	Non-zero variance components	Cluster/kernel	Design	σ_g^2[†]	Method
Scenario 1: *Selection under different sample sizes*					
$n = 20,\ 50,\ 100$; simulated count data	$l = 1,\ 2,$ $3,\ 4,\ 5$	genus; $\boldsymbol{K}_W$	longitudinal; cross-sectional	$1,\ 5,$ $25,\ 100$	VC-lasso group-lasso
Scenario 2: *Selection under different number of non-zero variance components*					
$n = 50$; simulated count data	(i) $l = 20,\ 30$; (ii) $l = 1,\ 2,$ $3,\ 4,\ 5$; (iii) $l = 1,\ 2,$ $3, \dots,\ 15$;	genus; $\boldsymbol{K}_W$	longitudinal; cross-sectional	$1,\ 5,$ $25,\ 100$	VC-lasso group-lasso
Scenario 3: *Selection under different UniFrac distance kernels*					
$n = 50$; simulated count data	$l = 1,\ 2,$ $3,\ 4,\ 5$	genus; $\boldsymbol{K}_W$, $\boldsymbol{K}_{UW}$, $\boldsymbol{K}_{VAW}$, $\boldsymbol{K}_0$, $\boldsymbol{K}_{0.5}$	longitudinal; cross-sectional	$1,\ 5,$ $25,\ 100$	VC-lasso group-lasso
Scenario 4: *Selection under fixed effect model*					
$n = 50$; simulated count data	$l = 20,\ 30$; $l = 1,\ 2,$ $3,\ 4,\ 5$; $l = 1,\ 2,$ $3, \dots,\ 15$;	genus; $\boldsymbol{K}_W$	cross-sectional	$1,\ 5,$ $25,\ 100$	VC-lasso group-lasso

Throughout simulations, $\sigma_e^2 = 1$, $\beta_1 = \beta_2 = 0.1$. We use $\sigma_d^2 = 0.6$ and 3 repeated measurements in the longitudinal design. We use $\sigma_d^2 = 0$ for the cross-sectional design. Group-lasso is performed only in the cross-sectional design.

[†] *The non-zero variance components are assumed to have equal effect strength in each simulation setting.*

3. SIMULATION

In this section, we conduct simulation studies to evaluate the variable selection and prediction performance of VC-lasso and compare the results with the conventional method group-lasso as implemented in the `gglasso` package (Yang and Zou, 2015). Phenotypes are simulated based on one real pulmonary microbiome dataset and one simulated longitudinal microbiome dataset. We first describe real and simulated microbiome abundance data, phylogenetic tree, and then detail our four phenotype simulation schemes (**Table 1**).

The real pulmonary microbiome data has been discussed in Twigg III et al. (2016). Thirty individuals were recruited. During up to three-years follow-up, lung functions and microbiome composition were measured 2–4 times for each individual. The longitudinal microbiome taxonomic data is summarized as 2,964 OTUs with a phylogenetic tree (Twigg III et al., 2016). Longitudinal microbiome abundance data is generated by a Zero-Inflated Beta Random Effect model using `R` package `ZIBR` in Supplementary Material section 2 (Chen and Li, 2016). For cross-sectional design, we generate taxonomic data using a Dirichlet-Multinomial (DM) model (Chen et al., 2012a). Simulation parameters, such as proportion of each OTU and the overall dispersion, are estimated from our real pulmonary microbiome abundance data.

Given simulated microbiome count data and taxonomic information, we classify 2,353 of 2,964 OTUs to 30 genera (taxa clusters) and the remaining 611 of 2,964 OTUs are grouped into the 31st cluster named *other* (**Table 2**). As described in Supplementary Material section 1, UniFrac distance matrices ($\boldsymbol{D}$) of the 31 clusters are computed and converted to kernel matrices as

$$\boldsymbol{K} = -\frac{1}{2}(\boldsymbol{I} - \frac{\boldsymbol{11}'}{n})\boldsymbol{D}^2(\boldsymbol{I} - \frac{\boldsymbol{11}'}{n}) \qquad (9)$$

followed by a positive definiteness correction (Chen and Li, 2013; Zhao et al., 2015). All of the microbiome kernel matrices $\boldsymbol{K}$ are scaled to have unit Frobenius norm.

Phenotypes are simulated based on the following scenarios.

3.1. Scenario 1: Selection Under Different Sample Size

Longitudinal and cross-sectional responses are generated by

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2,\, \sigma_d^2\boldsymbol{ZZ}' + \sum_{l=1}^{L}\sigma_{gl}^2\boldsymbol{K}_l + \sigma_e^2\boldsymbol{I}), \qquad (10)$$

where $\sigma_{gl}^2 > 0$ for $l = 1, \ldots, 5$ and $\sigma_{gl}^2 = 0$ otherwise. The total number of variance components for microbiome clusters is $L = 31$. The true model has five non-zero variance components including *Anaerococcus*, *Atopobium*, *Actinomyces*, *Campylobacter*, and *Capnocytophaga*. We compare the selection performance at three sample sizes: $n = 20,\ 50,\ 100$. For cross-sectional design, responses are simulated by setting $\sigma_d^2 = 0$.

TABLE 2 | List of 31 Genera.

	Genus	Phylum	No of OTU	Mean Reads
1	*Actinomyces*	*Actinobacteria*	150	230.59
2	*Anaerococcus*	*Firmicutes*	17	2.90
3	*Atopobium*	*Actinobacteria*	22	40.83
4	*Campylobacter*	*Proteobacteria*	31	51.05
5	*Capnocytophaga*	*Bacteroidetes*	31	70.81
6	*Catonella*	*Firmicutes*	22	40.09
7	*Corynebacterium*	*Actinobacteria*	47	12.22
8	*Flavobacterium*	*Bacteroidetes*	25	5.08
9	*Fusobacterium*	*Fusobacteria*	55	174.29
10	*Gemella*	*Firmicutes*	17	72.11
11	*Lactobacillus*	*Firmicutes*	33	141.10
12	*Leptotrichia*	*Fusobacteria*	15	12.40
13	*Megasphaera*	*Firmicutes*	14	36.99
14	*Methylobacterium*	*Proteobacteria*	11	2.88
15	*Neisseria*	*Proteobacteria*	18	109.61
16	*OD1_genera_incertae_sedis*	*OD1*	75	0.92
17	*Parvimonas*	*Firmicutes*	20	76.46
18	*Peptoniphilus*	*Firmicutes*	11	1.16
19	*Porphyromonas*	*Bacteroidetes*	42	134.41
20	*Prevotella*	*Bacteroidetes*	304	833.35
21	*Rothia*	*Actinobacteria*	16	49.83
22	*Selenomonas*	*Firmicutes*	50	16.16
23	*Sneathia*	*Fusobacteria*	12	37.09
24	*Sphingomonas*	*Proteobacteria*	14	0.61
25	*SR1_genera_incertae_sedis*	*SR1*	17	5.95
26	*Streptococcus*	*Firmicutes*	66	1,107.81
27	*TM7_genera_incertae_sedis*	*TM7*	61	40.54
28	*Treponema*	*Spirochaetes*	60	51.62
29	*Unclassified*	*Unclassified*†	1,068	258.65
30	*Veillonella*	*Firmicutes*	29	370.85
31	*Others*	*Others*	611	1,009.88

Summary of phylum information, the number of OTUs, and the average abundance (across sample and time points) within each genus from the pulmonary microbiome dataset are shown.

†The genus unclassified may belong to phylum unclassified or other 12 phyla.

3.2. Scenario 2: Selection Under Different Numbers of Non-zero Variance Components

The sample size is fixed at $n = 50$ in this scenario. Responses are generated by model (Equation 10) with different numbers of non-zero variance components. In Supplementary Material section 5, VC-lasso is evaluated when the number of variance components in the model is large.

(1) 2 non-zero variance components: $\sigma_{g_{20}}^2 > 0$, $\sigma_{g_{30}}^2 > 0$, and $\sigma_{g_l}^2 = 0$ otherwise. Two associated genera are *prevolleta* and *veillonella*.

(2) 5 non-zero variance components: $\sigma_{g_l}^2 > 0$ for $l = 1, 2, \ldots,\ 5$ and $\sigma_{g_l}^2 = 0$ otherwise. Associated clusters are *Anaerococcus*, *Atopobium*, *Actinomyces*, *Campylobacter*, and *Capnocytophaga*.

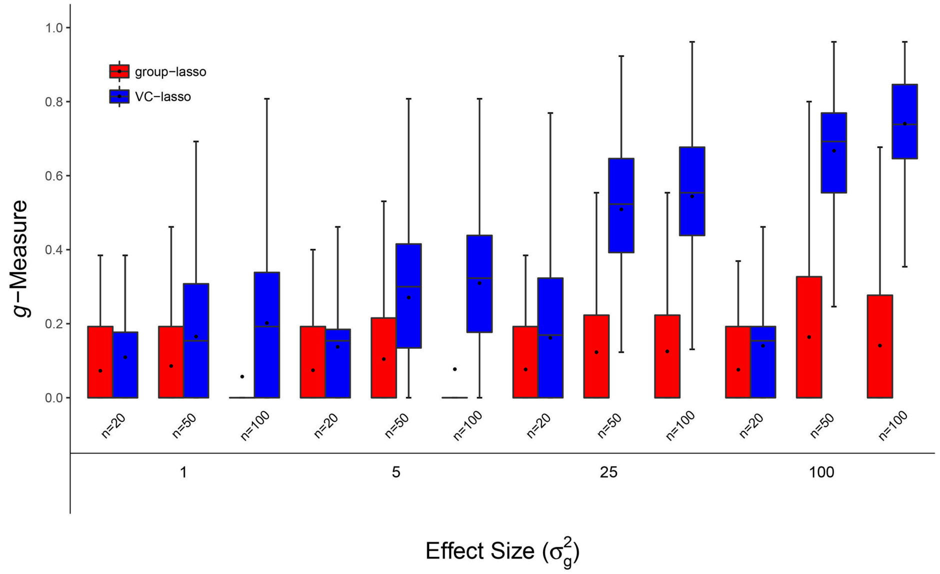


FIGURE 1 | Scenario 1: Estimated ***g***-Measure of both VC-lasso and group-lasso under different sample sizes for models with 5 non-zero variance components in a cross-sectional design. Three sample sizes, $n = 20, 50, 100$, are compared and $\sigma_d^2 = 0$.

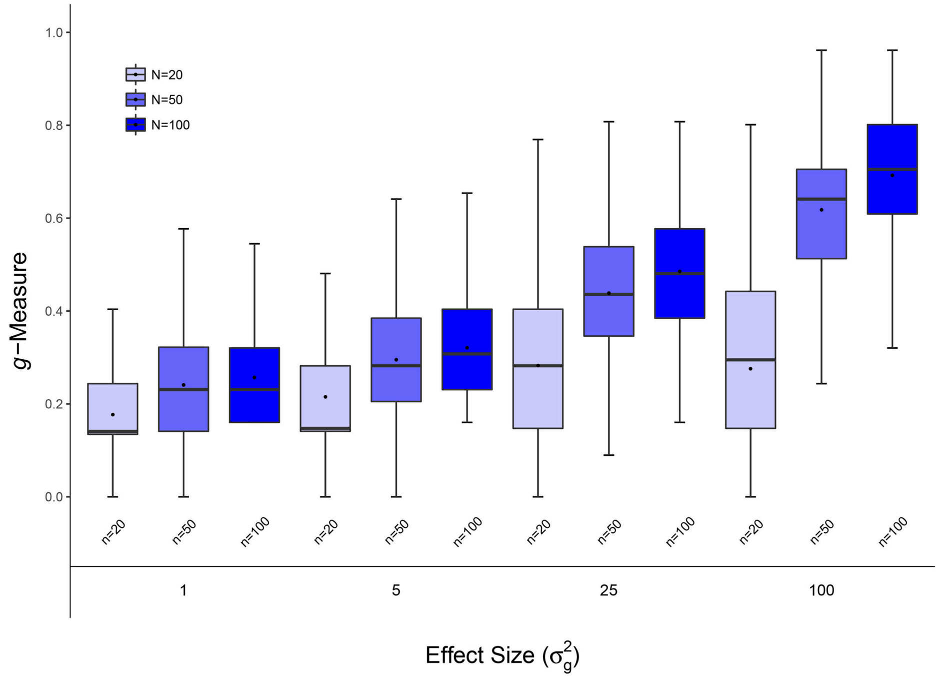


FIGURE 2 | Scenario 1: Estimated ***g***-Measure of VC-lasso under different sample sizes for models with 5 non-zero variance components in a longitudinal design. Three sample sizes, $n = 20, 50, 100$, are compared and $\sigma_d^2 = 0.6$.

(3) 15 non-zero variance components: $\sigma_{gl}^2 > 0$ for $l = 1, 2, \ldots,$ 15 and $\sigma_{gl}^2 = 0$ otherwise. Associated clusters, including *Actinomyces*, *Anaerococcus*, ..., and *Neisseria* are listed in **Table 2**.

3.3. Scenario 3: Selection Under Different UniFrac Distance Kernels

The sample size is fixed at $n = 50$ with 5 non-zero variance components. We compare the selection performance

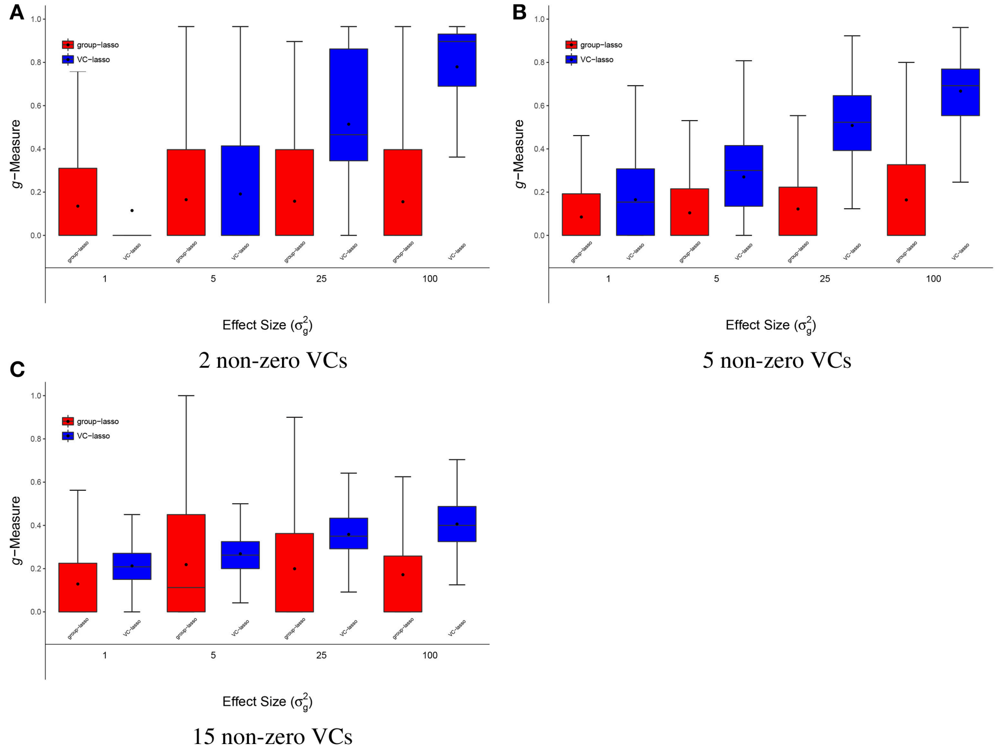


FIGURE 3 | Scenario 2: Estimated ***g***-Measure of both VC-lasso and group-lasso under different number of non-zero variance components in a cross-sectional design. The number of non-zero variance components (VCs) are set to 2 **(A)**, 5 **(B)**, 15 **(C)**, sample size is $n = 50$, and $\sigma_d^2 = 0$.

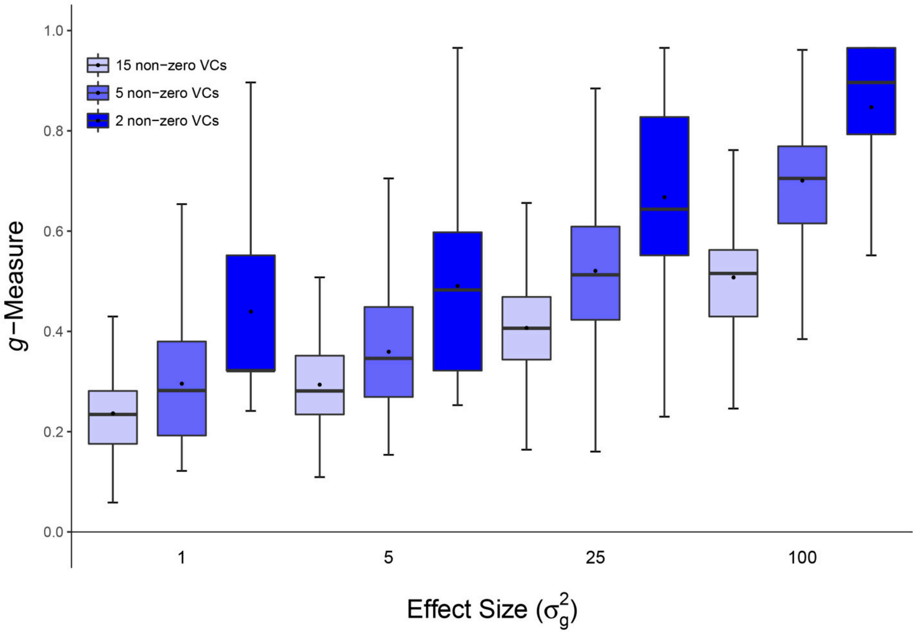


FIGURE 4 | Scenario 2: Estimated ***g***-Measure of VC-lasso under different number of non-zero variance components in a longitudinal design. Three different numbers of non-zero variance components (VCs), $2,\ 5,\ 15$, are shown, sample size is set to $n = 50$ and $\sigma_d^2 = 0.6$.

using kernels defined by 5 different distance measures: variance adjusted weighted UniFrac distance ($\boldsymbol{K}_{VAW}$) Chang et al., 2011), generalized UniFrac distance ($\boldsymbol{K}_0$, $\boldsymbol{K}_{0.5}$) (Chen et al., 2012a), unweighted UniFrac distance ($\boldsymbol{K}_{UW}$) (Lozupone and Knight, 2005), and weighted UniFrac distance ($\boldsymbol{K}_W$) (Lozupone et al., 2007).

3.4. Scenario 4: Selection Under Fixed Effect Model

We again use the sample size $n = 50$ and vary the number of clusters containing signal. Responses are simulated by a fixed effect model

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{G}_1^*\gamma_1 + \boldsymbol{G}_2^*\gamma_2 + \dots + \boldsymbol{G}_u^*\gamma_u,\ \sigma_e^2\boldsymbol{I}), \tag{11}$$

where $\boldsymbol{G}_1^*$, $\boldsymbol{G}_2^*, \dots, \boldsymbol{G}_u^*$ are OTU count matrices of different clusters scaled by their sample maximum. u is the total number of clusters with effects that ranges from 2 to 15. Fixed effect vector $\boldsymbol{\gamma}_l$ for cluster l are generated from $\boldsymbol{\gamma}_l \sim \mathcal{N}(0, \sigma_{gl}^2\boldsymbol{I})$ and are fixed for each simulation replicate.

We applied VC-lasso to scenarios 1-3 using both longitudinal and cross-sectional designs. Scenario 4 is performed using a cross-sectional design only. We compare our approach with group-lasso (R package gglasso) in all 4 scenarios for cross-sectional design because the gglasso package cannot handle longitudinal data.

We set the within-individual variance $\sigma_e^2 = 1$ throughout simulations. The between individual variance of random intercept is set to $\sigma_d^2 = 0.6$ for longitudinal design and $\sigma_d^2 = 0$ otherwise (Twigg III et al., 2016). The effect strength is set to $\sigma_g^2 = 1, 5, 25, 100$ (Chen et al., 2015a). We set the non-zero variance components to have the same effect strength under each setting, therefore omit subscript l. Two covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are generated from the standard normal distribution and effect sizes are set to $\beta_1 = \beta_2 = 0.1$. 1000 Monte Carlo simulation replicates

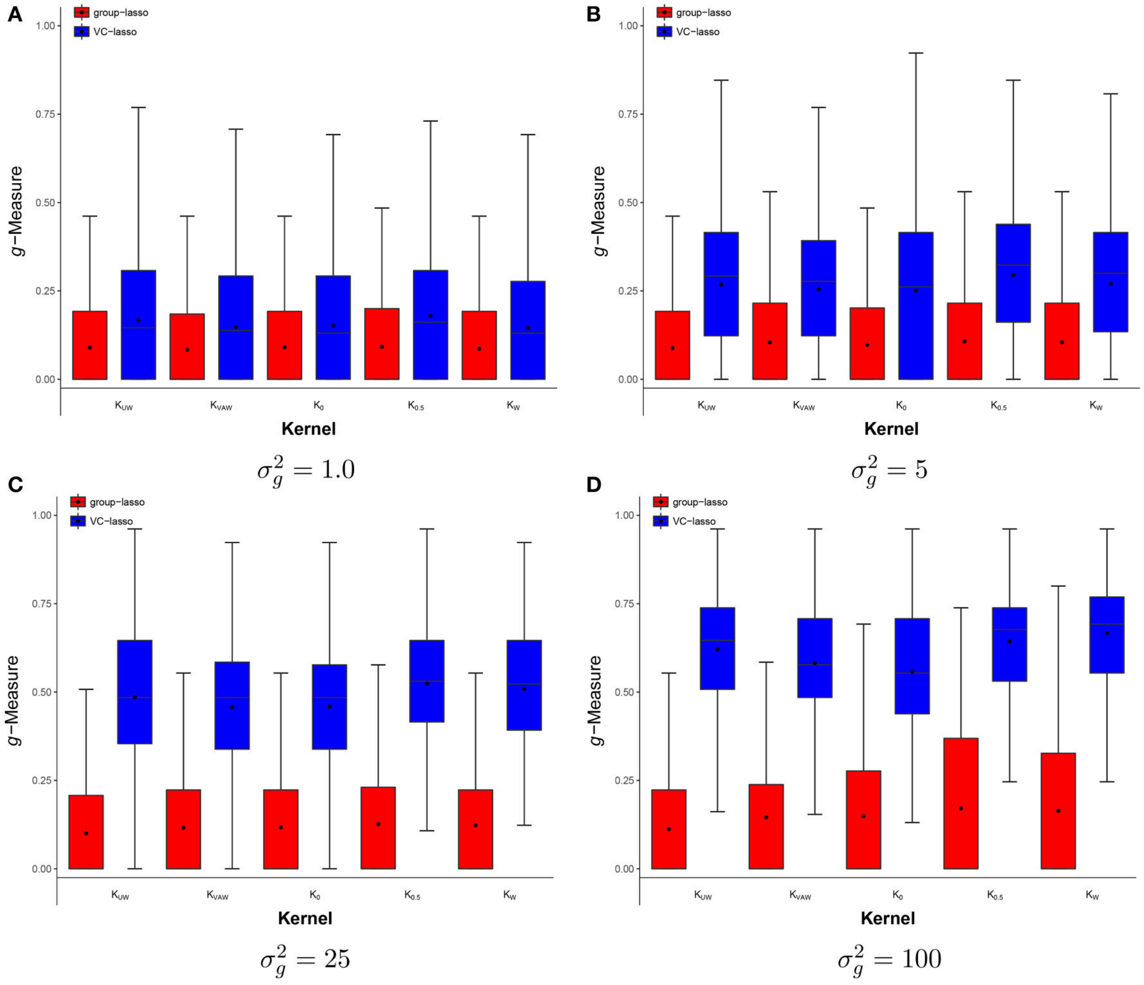


FIGURE 5 | Scenario 3: Estimated ***g***-Measure of both VC-lasso and group-lasso under different UniFrac distance kernels in a cross-sectional design. Five different kernels, K_{UW}, K_{VAW}, K_0, $K_{0.5}$ and K_W, and two methods, VC-lasso and group-lasso, are displayed in a cross-sectional design. Four effect strengths, 1 **(A)**, 5 **(B)**, 25 **(C)**, and 100 **(D)** are shown. There are 5 non-zero variance components, sample size is $n = 50$, and $\sigma_d^2 = 0$.

are generated. We split each dataset to training (80%) and testing (20%). Five-fold cross-validation is performed in training set to estimate the optimal λ^*. Selection performance is evaluated and reported by applying λ^* to the testing set.

4. RESULTS

4.1. Analysis of Simulated Data

The simulation results are summarized in **Figures 1–9** including variable selection performance under different sample sizes (**Figures 1**, **2**), different numbers of non-zero variance components (**Figures 3**, **4**), and different UniFrac distance measures (**Figures 5**, **6**) for both cross-sectional and longitudinal designs. Comparisons between VC-lasso and group-lasso are shown in all cross-sectional simulation studies.

The trajectories of g-Measure versus tuning parameter λ from cross-validation is presented in **Figure 9**. g-Measures remain stable or slightly decrease as λ getting larger under moderate effect size when $\sigma_g^2 = 1$ and 5. It starts to decrease when λ is greater than 0.6. **Figure 9** suggests that the trajectories of tuning criteria is generally consistent across sample sizes, effect sizes, and study designs.

4.1.1. Scenario 1: Selection Under Different Sample Sizes

Figures 1, **2**, **8A** display performance of selection (g-Measure) and prediction (area under the receiver operating characteristic curve, AUROC). In **Figure 1**, we compare VC-lasso (blue bar) and group-lasso (red bar) using cross-sectional design. In **Figure 2**, we compare the g-Measure of VC-lasso under different sample sizes using a longitudinal design. For both cross-sectional and longitudinal designs, g-Measure of VC-lasso boosts with increased sample size and effect sizes. Except for the third quartile of g-Measure over 1,000 replicates for sample size, $n = 20$, VC-lasso always outperforms group-lasso in this scenario.

Area under receiver operating characteristic (AUROC) is used to evaluate the prediction performance (**Figure 8A**) when effect size is fixed at $\sigma_g^2 = 25$. Larger AUROC represents better prediction ability. For VC-lasso, AUROC increases with sample size under cross-sectional design. For longitudinal study, $n = 50$ has similar AUROC with $n = 100$, which indicates the optimal prediction we can receive under this simulation setting. The AUROC of group-lasso (red bar) is similar under different sample sizes and shows no advantages compared to the VC-lasso.

4.1.2. Scenario 2: Selection Under Different Number of Non-zero Variance Components

Figures 3, **4**, **8B** show simulation results for the selection under different number of non-zero variance components. Specifically, **Figure 3** shows g-Measure for both VC-lasso and group-lasso in a cross-sectional design, while **Figure 4** presents g-Measure for VC-lasso in a longitudinal design.

In **Figures 3**, **4**, the performance of VC-lasso selection improves when effect size increases. For a model with 2 non-zero variance components, the true discovery rate (TDR, or sensitivity) is either 0, 0.5 or 1.0, which lead to a large variation of the g-Measure (**Figure 3A**). As more non-zero variance

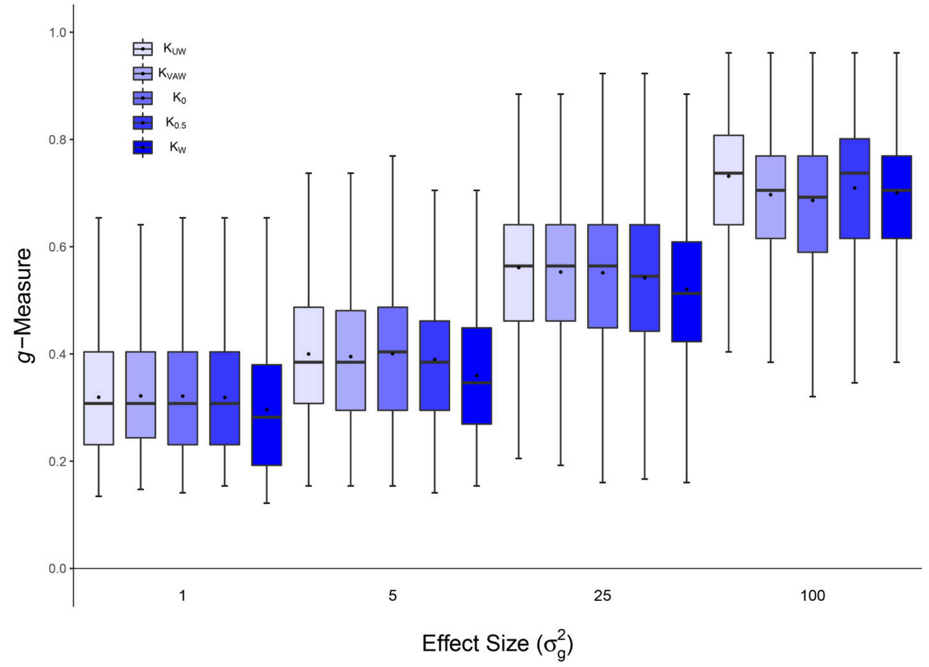
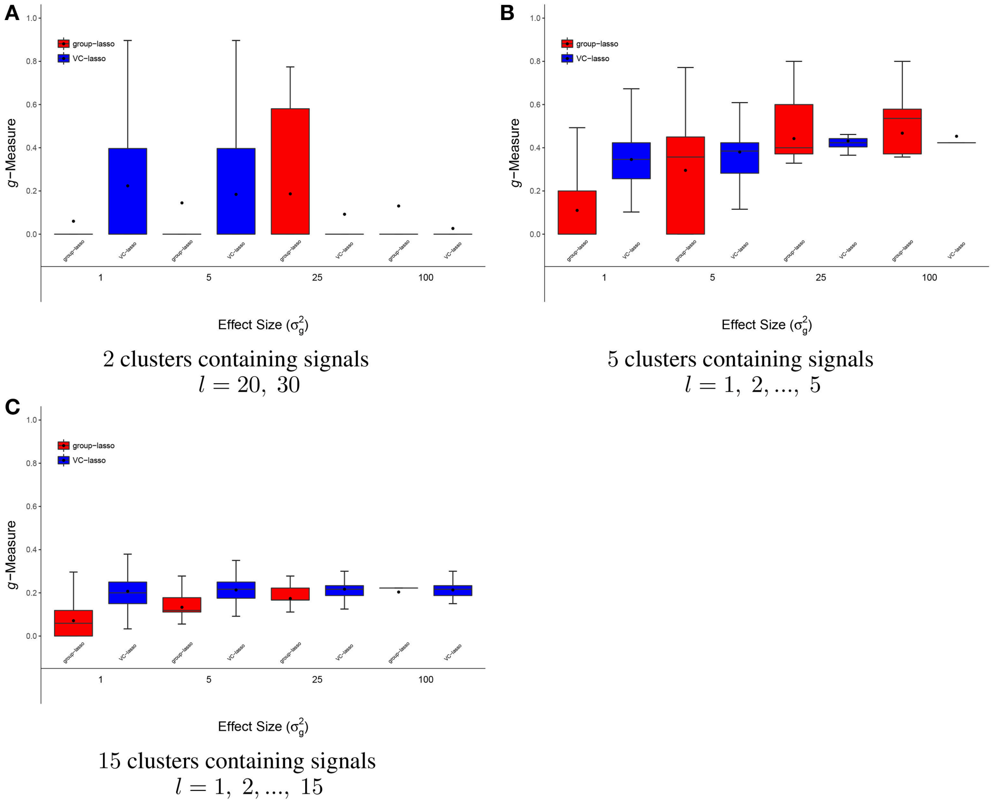


FIGURE 6 | Scenario 3: Estimated $\boldsymbol{g}$-Measure of VC-lasso under different UniFrac distance kernels in a longitudinal design.Five different kernels, K_{UW}, K_{VAW}, K_0, $K_{0.5}$ and K_W, are compared. There are 5 non-zero variance components. Sample size is $n = 50$ and $\sigma_d^2 = 0.6$.

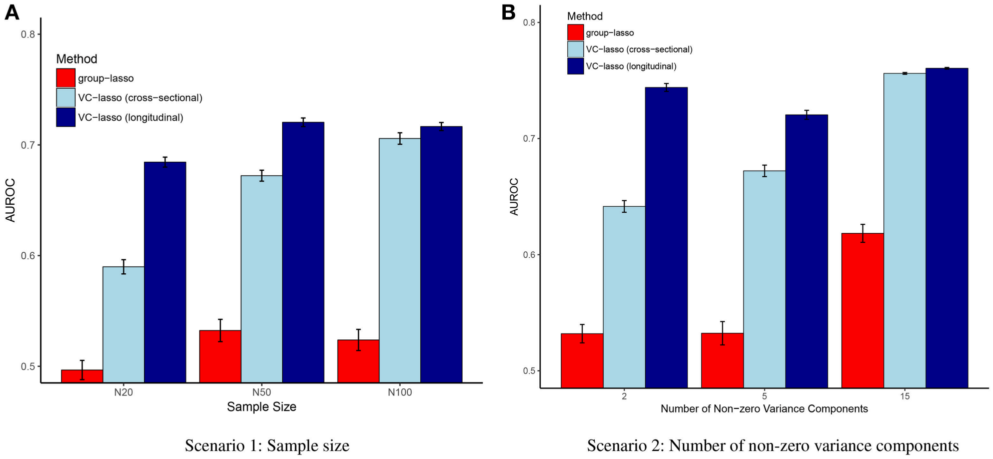


2 clusters containing signals
$l = 20,\ 30$

5 clusters containing signals
$l = 1,\ 2, \dots,\ 5$

15 clusters containing signals
$l = 1,\ 2, \dots,\ 15$

FIGURE 7 | Scenario 4: Estimated ***g***-Measure of VC-lasso and group-lasso under fixed effect model in a cross-sectional design. There are 2 **(A)**, 5 **(B)**, 15 **(C)** clusters with signals. Sample size is $n = 50$ and $\sigma_d^2 = 0$.

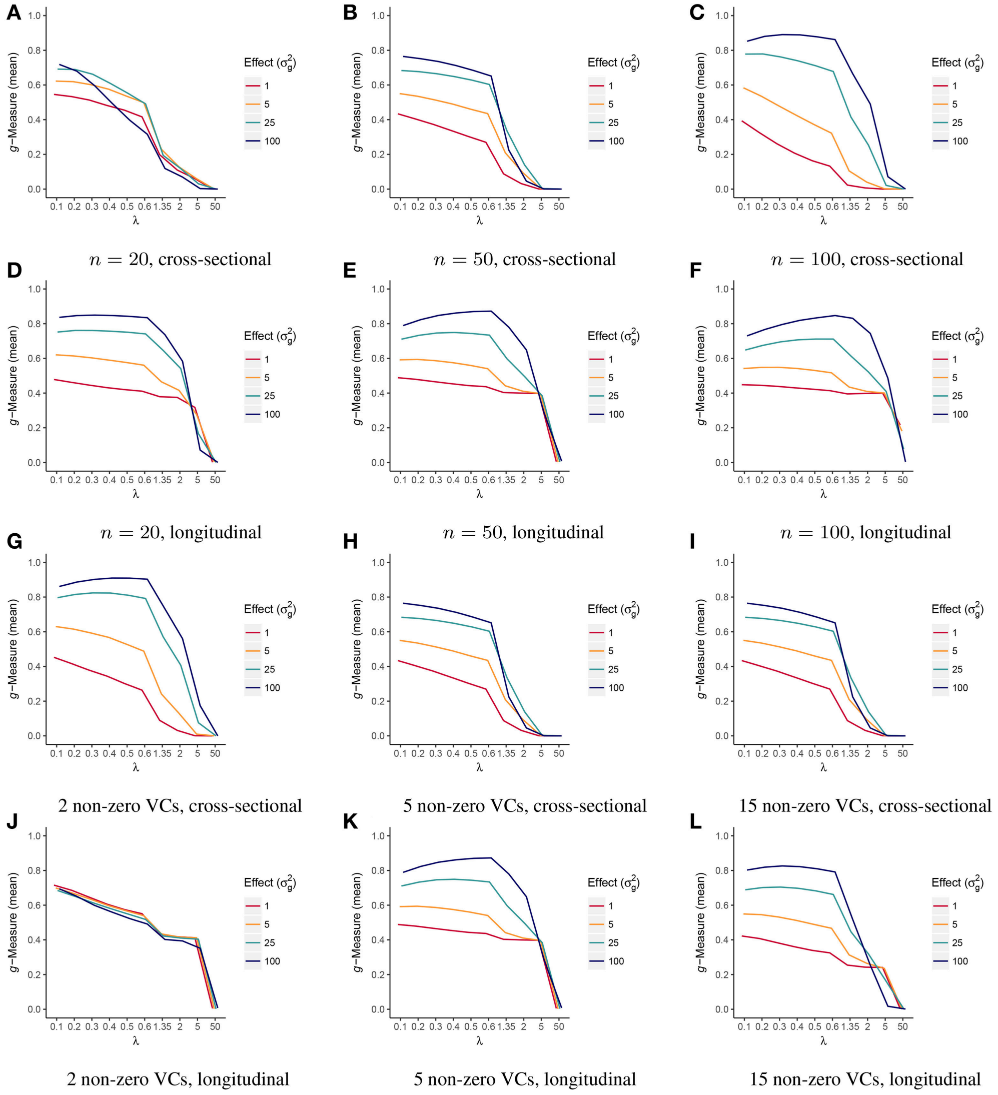


Scenario 1: Sample size

Scenario 2: Number of non-zero variance components

FIGURE 8 | Scenario 1 & 2: AUROC. The AUROC is presented as the mean ± 95% confidence interval based on 1,000 simulation replicates for each simulation scenario when $\sigma_g^2 = 25$. **(A)** Scenario 1; **(B)** Scenario 2.

components are included, in **Figure 3B,C** the trajectory of *g*-Measures becomes smoother. The *g*-Measures of VC-lasso are higher than the group-lasso in most simulation settings except that group-lasso has larger third quartile when $\sigma_g^2 = 1$ in **Figure 3A** and $\sigma_g^2 = 5$ in **Figure 3C**. As shown in **Figure 8B**, VC-lasso has a better prediction ability with an increased number of non-zero variance components. Compared with our method, group-lasso is uncompetitive in predictive ability.



n = 20, cross-sectional

n = 50, cross-sectional

n = 100, cross-sectional

n = 20, longitudinal

n = 50, longitudinal

n = 100, longitudinal

2 non-zero VCs, cross-sectional

5 non-zero VCs, cross-sectional

15 non-zero VCs, cross-sectional

2 non-zero VCs, longitudinal

5 non-zero VCs, longitudinal

15 non-zero VCs, longitudinal

FIGURE 9 | Trajectories of estimated ***g***-Measure as a function of tuning parameter **λ** (scenario 1 & 2). Estimated *g*-Measure is displayed as the mean of 5-fold cross-validation under sample sizes, $n = 20$ **(A,D)**, 50 **(B,E)**, 100 **(C,F)**, or 2 **(G,J)**, 5 **(H,K)**, 15 **(I,L)** non-zero variance components (VCs) in both cross-sectional and longitudinal designs.

4.1.3. Scenario 3: Selection Under Different UniFrac Distance Kernels

We compare the *g*-Measure of five different kernels in **Figures 5**, **6** for the cross-sectional and longitudinal design, respectively. Using longitudinal simulated data, the box-plot of *g*-Measure shows that the five kernels have similar performance except that the K_W has the lowest third quartile and K_0 has the lowest first quartile when σ_g^2 is large. Under the same effect strength (σ_g^2) in the cross-sectional design (**Figure 5**), the *g*-Measure of five kernels are almost identical except that K_0 has slightly smaller *g*-Measure and wider range than other kernels. For example, K_0 has the lowest first quartile in **Figures 5B**. This suggests that the kernels computed from different UniFrac distance play a minor part in the selection performance and our method is superior to group-lasso regardless of kernel types.

4.1.4. Scenario 4: Selection Under Fixed Effect Model

Figure 7 has a distinctive pattern from the above scenarios. For the case that only two microbiome clusters contain signals (*Prevotella* and *Veillonella*), both methods do not perform well (**Figure 7A**). In **Figures 7B,C**, *g*-Measures for both methods improve with increased effect sizes and VC-lasso outplays group-lasso with $\sigma_g^2 = 1$. For $\sigma_g^2 = 5, 25$, average and median *g*-Measure of VC-lasso across simulation replicates outperform group-lasso. Besides, we notice that the range of *g*-Measure for VC-lasso becomes smaller as signal strengths increase, suggesting the prediction performance stabilizes as the association with the

TABLE 3 | Analysis of Forced expiratory volume in one second (FEV1) at genus level in the real pulmonary microbiome cohort using variance component lasso selection (VC-lasso) and exact tests.

	VC-lasso			Exact tests		
	Rank	**Genus**	**Phylum info**	**eRLRT**	**eLRT**	**eScore**
Baseline	1	*Corynebacterium*	*Actinobacteria*	0.28	0.30	0.30
	2	*TM7_genera_incertae_sedis*	*TM7*	1.00	1.00	1.00
	3	*Anaerococcus*	*Firmicutes*	0.06	0.06	0.07
	4	*Neisseria*	*Proteobacteria*	1.00	1.00	1.00
	5	*Treponema*	*Spirochaetes*	0.13	0.14	0.14
Longitudinal	1	*Corynebacterium*	*Actinobacteria*	1.00	–	1.00
	2	*Actinomyces*	*Actinobacteria*	0.00	–	0.01
	3	*Prevotella*	*Bacteroidetes*	0.01	–	0.01
	4	*TM7_genera_incertae_sedis*	*TM7*	1.00	–	1.00
	5	*Porphyromonas*	*Bacteroidetes*	0.00	–	0.00
	6	*Megasphaera*	*Firmicutes*	0.06	–	0.06

The phylum information is provided for selected genera. Tuning parameter λ∗ for baseline and longitudinal data is set to 0.01 and 0.2, respectively. Rank represents the order of genera that appear in the solution path. Results of eLRT are omitted as it is equivalent to eRLRT in a longitudinal design.

TABLE 4 | Analysis of forced expiratory flow (FEF) at genus level in the real pulmonary microbiome cohort using variance component lasso selection (VC-lasso) and exact tests.

	VC-lasso			Exact tests		
	Rank	**Genus**	**Phylum info**	**eRLRT**	**eLRT**	**eScore**
Baseline		–	–	–	–	–
Longitudinal	1	*Methylobacterium*	*Proteobacteria*	1.00	–	1.00
	4	*Prevotella*	*Bacteroidetes*	<0.01	–	<0.01
	2	*Rothia*	*Actinobacteria*	0.01	–	0.03
	3	*Campylobacter*	*Proteobacteria*	0.03	–	0.03
	5	*TM7_genera_incertae_sedis*	*TM7*	0.00	–	0.01
	6	*Corynebacterium*	*Actinobacteria*	0.32	–	0.31

The phylum information is provided for selected genus. Tuning parameter $\lambda^* = 0.035$ for longitudinal data. Rank represents the order of genera that appear in the solution path. No genus is chosen using baseline data only. Results of eLRT are omitted in longitudinal design as it is equivalent to eRLRT.

outcome increases. In general, VC-lasso has a distinctively better selection performance even when model is misspecified.

4.2. Application to Longitudinal Pulmonary Microbiome Data

We apply VC-lasso to a longitudinal dataset of pulmonary microbiome study. Bronchoalveolar lavage (BAL) fluid were collected for microbiome profiling. The inclusion criterion for this cohort were: (1) HIV infection and (2) CD4 count less than 500 $cells/mm^3$ before HAART (Twigg III et al., 2016). Two most common pulmonary function tests were performed repeatedly: spirometry and diffusing capacity for carbon monoxide. In this report we focus on spirometry measures. Spirometry is to measure the lung volume and how well the lung exhales, such as average forced expiratory flow (FEF) and forced expiratory volume in 1s (FEV1). Both spirometry and diffusing capacity were evaluated as percent predicted values as pulmonary function tests are usually interpreted by comparing the patient's value to predicted value of the healthy subject with similar age, height and ethnicity (Twigg III et al., 2016).

Twigg III et al. (2016) compared microbiome abundance differences at overall community level between (1) uninfected and baseline; (2) uninfected and 1 year after treatment; and (3) uninfected and 3 year treated subjects. They suggest that the lung microbiome in healthy HIV-infected individuals with preserved CD4 counts is similar to uninfected individuals. Among individuals with more advanced disease, there is an altered alveolar microbiome characterized by a loss of richness and evenness (alpha diversity). This alteration might impact pulmonary complications (often characterized by the measure of lung functions) in HIV-infected patients on antiretroviral therapy (ART). In this application, we therefore aim to identify microbiome genera associated with pulmonary function in both longitudinal and baseline studies. Ethnicity, gender, smoking history, CD4 count, and HIV viral load are included as the covariates. Missing covariates are imputed

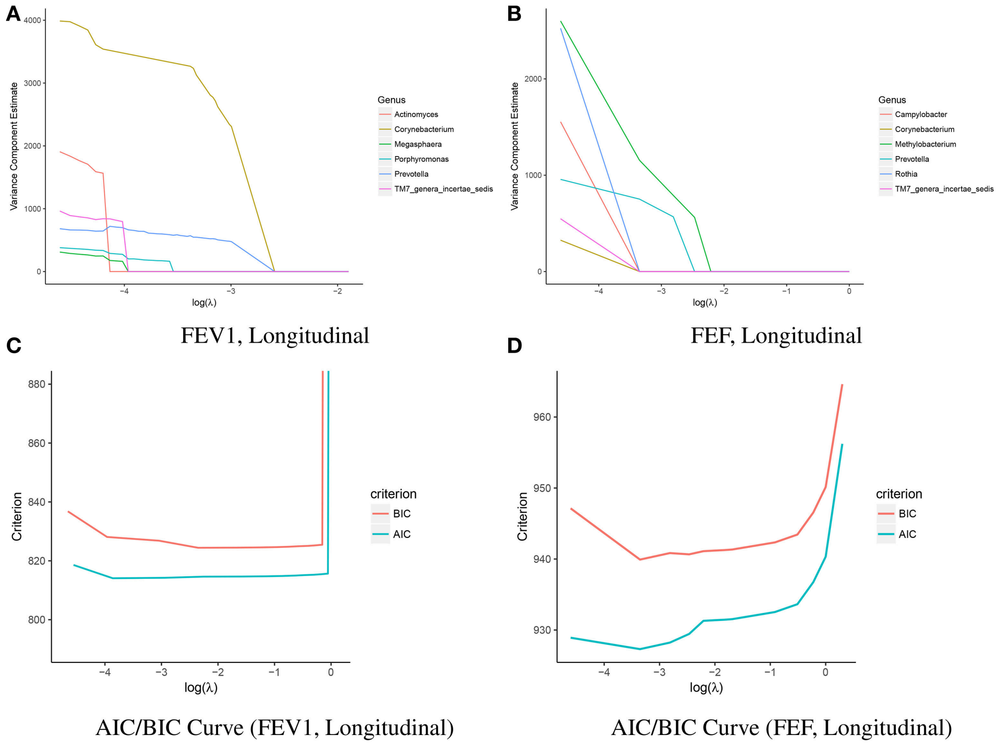


FIGURE 10 | Solution path and AIC/BIC curve of the VC-lasso method in the analysis of 31 genera and the pulmonary function. The solution paths with penalty parameter are presented for FEV1 **(A)** and FEF **(B)** in longitudinal study (upper panel). AIC/BIC curves as a function of tuning parameter for FEV1 **(C)** and FEF **(D)** are shown in the lower panel.

by their mean. Penalized variance component selection is performed among all 31 genera. Due to limited sample sizes, we choose the optimal tuning parameter λ^* by AIC and BIC.

Tables 3, **4** show selected genera with their phylum information and the corresponding *p*-values from exact tests, i.e., score test (eScore), likelihood ratio test (eLRT), and restricted likelihood ratio test (eRLRT) (Zhai et al., 2017b). The genera are ranked in the order they appear in the solution path (**Figures 10A,B**). VC-lasso selects 6 genera associated with FEV1 using longitudinal data and $\lambda^* = 0.2$ (**Table 3** and **Figure 10C**). Three out of six selected genera have eRLRT *p*-values < 0.05 (**Table 3**), including *Actinomyces* ($p < 0.01$), *Prevotella* ($p = 0.01$), and *Porphyromonas* ($p < 0.01$). Using baseline data, we identify five genera associated with FEV1, among which *Corynebacterium* and *TM7 genera incertae sedis* are also selected by using longitudinal data. Several selected genera received insufficient attention in HIV-infected populations previously, for example, *Anaerococcus* and *Megasphaera*. Studies have shown that *Anaerococcus* became more abundant in children with asthma after azithromycin treatment (Slater et al., 2013; Riiser, 2015) and *Megasphaera* has higher relative abundance in smoking population (Segal et al., 2014). However, none of them has been reported in HIV infected pulmonary microbiome (Rogers et al., 2004; Chen et al., 2007; Twigg III et al., 2016).

For variance component selection on FEF (**Table 4**), VC-lasso selects 6 genera in total using longitudinal data with $\lambda^* = 0.035$. Considering the exact test results (eRLRT and eScore), four of them show significant association with FEF (*p*-value < 0.05), i.e., *Prevotella*, *Rothia*, *Campylobacter* and *TM7_genera_incertae_sedis*. Twigg III et al. (2016) reported that HIV-positive BAL samples contained an increased abundance of *Prevotella* after 1-year HAART treatment while significantly decreased abundances during 3 years of treatment. *Campylobacter* is another noteworthy genus that has significant association with inflammation markers of HIV-infected population (Iwai et al., 2014). Additionally, significantly increased abundance of *Rothia* and *TM7_genera_incertae_sedis* in oral wash microbiome has been reported in HAART treatment group (Iwai et al., 2012; Beck et al., 2015). In conclusion, VC-lasso provides innovative association evidence between fine level pulmonary microbiome clusters with lung function phenotypes. Our report is a hypothesis generation procedure. Association results need to be further validated in a separate population or by laboratory experiments.

5. DISCUSSION

In this paper, we propose the variance component selection scheme VC-lasso for sparse and high-dimensional taxonomic data analysis. To reduce the dimensionality, we first aggregate the dispersed individual OTUs to clusters at higher phylogenetic level, such as genus, family, or phylum. By translating the phylogenetic distance information to kernel matrices, we treat the aggregated taxonomic clusters as multiple random effects in a variance component model. Then, VC-lasso is performed for parsimonious variable selection of variance components. The MM algorithm with lasso penalization derived in Algorithm 1 for parameter estimation extremely simple and computationally efficient for variance component estimation. The group-lasso as a comparison can also be used for the microbiome cluster selection and incorporating higher phylogenetic group information (Yuan and Lin, 2006; Garcia et al., 2013; Yang and Zou, 2015). However, group-lasso suffers from the high-dimensionality and sparsity of OTUs within clusters. And group-lasso is not easy to accommodate phylogenic information. Beyond that, our novel approach VC-lasso can be applied to longitudinal designs. In such cases, we do not penalize the variance component that contains the repeat measurement correlation. Software and detailed documentation are freely available at https://github.com/JingZhai63/VCselection.

The VC-lasso is not limited to random intercept model for longitudinal studies. More complex random effect models, such as random intercept and random slope model, can also be used. More generally, the extension of our method to multivariate responses is expected to have better prediction performances. In the precision medicine era, with the rapid development of sequencing techniques and decreasing costs, the personal microbiome sequencing is already available to the consumer, e.g., American Gut (http://americangut.org/) and uBiome (https://ubiome.com/). Selection for higher-order interactions with random effect, such as microbiome and treatment regime interactions (Gopalakrishnan et al., 2017), will be a straightforward, yet interesting, implementation (Maity and Lin, 2011; Lin et al., 2016).

In practice, knowledge is needed about which taxonomy level should be aimed at to develop strategies for intervention. Considering multiple level taxonomic data, one can extend VC-lasso to include tree topologies (Wang and Zhao, 2016; Wang et al., 2017). For example, overlapping or subgroup VC-lasso can be developed by using both ℓ_1 and ℓ_2 regularizations (Jacob et al., 2009; Bien et al., 2013). Last but not the least, the variance components model requires specification of a kernel function or kernel matrix a priori, but it is often unclear which distance kernel to use in practice. To deal with the uncertainty, we can consider obtaining a composite kernel by utilizing a multiple kernel learning algorithm, such as a multi-kernel boosting algorithm (Xia and Hoi, 2013). In conclusion, with its competitive performance and many potential extensions, our variance components model with regularization, VC-lasso, is a powerful tool for mining the emerging microbiome data.

AUTHOR CONTRIBUTIONS

JZ implemented method and carried out data analysis. JZ wrote the manuscript with support from JK, HZ, and JJZ. HZ helped supervise the project. HT and KK provided pulmonary microbiome data. JJZ supervised the project.

ACKNOWLEDGMENTS

JJZ is supported by NIH grant K01DK106116 and Arizona Biomedical Research Commission (ABRC) grant. HZ is partially supported by NIH grants HG006139, GM105785, GM53275 and NSF grant DMS-1645093. HT and KK is supported by NIH grant UO1 HL121831 and UO1 HL098960. An allocation of computer time from the UA Research Computing High Performance Computing (HPC) and High Throughput Computing (HTC) at the University of Arizona is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00509/full#supplementary-material

REFERENCES

Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, eds E. Parzen, K. Tanabe, and G. Kitagawa (New York, NY: Springer), 199–213.

Beck, J. M., Schloss, P. D., Venkataraman, A., Twigg III, H., Jablonski, K. A., Bushman, F. D., et al. (2015). Multicenter comparison of lung and oral microbiomes of HIV-infected and HIV-uninfected individuals. *Am. J. Respirat. Crit. Care Med.* 192, 1335–1344. doi: 10.1164/rccm.201501-0128OC

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.* 41:1111. doi: 10.1214/13-AOS1096

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi: 10.2307/1942268

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

Chang, Q., Luan, Y., and Sun, F. (2011). Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* 12:118. doi: 10.1186/1471-2105-12-118

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. I., Kao, S. J., and Hsu, Y.-H. (2007). Pathophysiological mechanism of lung injury in patients with leptospirosis. *Pathology* 39, 339–344. doi: 10.1080/00313020701329740

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012a). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2012b). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14, 244–258. doi: 10.1093/biostatistics/kxs038

Chen, J., Just, A. C., Schwartz, J., Hou, L., Jafari, N., Sun, Z., et al. (2015a). CpGFilter: model-based CpG probe filtering with replicates for epigenome-wide association studies. *Bioinformatics* 32, 469–471. doi: 10.1093/bioinformatics/btv577

Chen, J., and Li, H. (2013). *Kernel Methods for Regression Analysis of Microbiome Compositional Data*. New York, NY: Springer.

Chen, L., Liu, H., Kocher, J.-P. A., Li, H., and Chen, J. (2015b). glmgraph: an R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* 31, 3991–3993. doi: 10.1093/bioinformatics/btv497

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C., Yu, W.-H., et al. (2010). The human oral microbiome. *J. Bacteriol.* 192, 5002–5017. doi: 10.1128/JB.00542-10

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., et al. (2011). Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS ONE* 6:e16384. doi: 10.1371/journal.pone.0016384

Fan, Y., and Li, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* 40:2043. doi: 10.1214/12-AOS1028

Garcia, T. P., Müller, S., Carroll, R. J., and Walzem, R. L. (2013). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* 30, 831–837. doi: 10.1093/bioinformatics/btt608

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

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., et al. (2017). Gut microbiome modulates response to anti–pd-1 immunotherapy in melanoma patients. *Science* 359, 97–103. doi: 10.1126/science.aan4236

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol.* 9, 244–253. doi: 10.1038/nrmicro2537

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hodkinson, B. P., and Grice, E. A. (2015). Next-generation sequencing: a review of technologies and tools for wound microbiome research. *Adv. Wound Care* 4, 50–58. doi: 10.1089/wound.2014.0542

Hui, F. K., Müller, S., and Welsh, A. (2017). Joint selection in mixed models using regularized PQL. *J. Am. Statist. Assoc.* 112, 1323–1333. doi: 10.1080/01621459.2016.1215989

Hunter, D. R., and Lange, K. (2004). A tutorial on MM algorithms. *Am. Statist.* 58, 30–37. doi: 10.1198/0003130042836

Hunter, D. R., and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* 3:1617. doi: 10.1214/009053605000000200

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* 67, 495–503. doi: 10.1111/j.1541-0420.2010.01463.x

Iwai, S., Fei, M., Huang, D., Fong, S., Subramanian, A., Grieco, K., et al. (2012). Oral and airway microbiota in HIV-infected pneumonia patients. *J. Clin. Microbiol.* 50, 2995–3002. doi: 10.1128/JCM.00278-12

Iwai, S., Huang, D., Fong, S., Jarlsberg, L. G., Worodria, W., Yoo, S., et al. (2014). The lung microbiome of Ugandan HIV-infected pneumonia patients is compositionally and functionally distinct from that of San Franciscan patients. *PLoS ONE* 9:e95726. doi: 10.1371/journal.pone.0095726

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). "Group lasso with overlap and graph lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, QC: ACM).

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* 34, 64–69. doi: 10.1038/nbt.3416

Lange, K. (2016). *MM Optimization Algorithms*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graphic. Statist.* 9, 1–20. doi: 10.1080/10618600.2000.10474858

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, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., et al. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* 72, 156–164. doi: 10.1111/biom.12368

Lozupone, C., Cota-Gomez, A., Palmer, B. E., Linderman, D. J., Charlson, E. S., Sodergren, E., et al. (2013). Widespread colonization of the lung by Tropheryma whipplei in HIV infection. *Am. J. Respirat. Crit. Care Med.* 187, 1110–1117. doi: 10.1164/rccm.201211-2145OC

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

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

Maity, A., and Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using garrote kernel machines. *Biometrics* 67, 1271–1284. doi: 10.1111/j.1541-0420.2011.01598.x

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141. doi: 10.1016/j.tig.2007.12.007

Morris, A., Beck, J. M., Schloss, P. D., Campbell, T. B., Crothers, K., Curtis, J. L., et al. (2013). Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *A. J. Respirat. Crit. Care Med.* 187, 1067–1075. doi: 10.1164/rccm.201210-1913OC

Riiser, A. (2015). The human microbiome, asthma, and allergy. *Allergy Asthma Clin. Immunol.* 11:35. doi: 10.1186/s13223-015-0102-0

Rogers, G., Carroll, M., Serisier, D., Hockey, P., Jones, G., and Bruce, K. (2004). Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16S ribosomal DNA terminal restriction fragment length polymorphism profiling. *J. Clin. Microbiol.* 42, 5176–5183. doi: 10.1128/JCM.42.11.5176-5183.2004

Schwarz, G. et al. (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464. doi: 10.1214/aos/1176344136

Segal, L. N., Rom, W. N., and Weiden, M. D. (2014). Lung microbiome for clinicians. new discoveries about bugs in healthy and diseased lungs. *Ann. Am. Thoracic Soc.* 11, 108–116. doi: 10.1513/AnnalsATS.201310-339FR

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Statist.* 10, 1019–1040. doi: 10.1214/16-AOAS928

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graphic. Statist.* 22, 231–245. doi: 10.1080/10618600.2012.681250

Slater, M., Rivett, D. W., Williams, L., Martin, M., Harrison, T., Sayers, I., et al. (2013). The impact of azithromycin therapy on the airway microbiota in asthma. *Thorax* 69, 673–674. doi: 10.1136/thoraxjnl-2013-204517

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

Twigg III, H. L., Knox, K. S., Zhou, J., Crothers, K. A., Nelson, D. E., Toh, E., et al. (2016). Effect of advanced HIV infection on the respiratory microbiome. *Am. J. Respirat. Crit. Care Med.* 194, 226–235. doi: 10.1164/rccm.201509-1875OC

Twigg III, H. L., Morris, A., Ghedin, E., Curtis, J. L., Huffnagle, G. B., Crothers, K., et al. (2013). Use of bronchoalveolar lavage to assess the respiratory microbiome: signal in the noise. *Lancet Respirat. Med.* 1, 354–356. doi: 10.1016/S2213-2600(13)70117-6

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14, 508–522. doi: 10.1038/nrmicro.2016.83

Wang, T., and Zhao, H. (2016). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Am. Statist. Associat.* 112, 1022–1031. doi: 10.1080/01621459.2016.1270213

Wang, T., and Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Statist.* 11, 771–791. doi: 10.1214/16-AOAS1017

Xia, H., and Hoi, S. C. (2013). Mkboost: A framework of multiple kernel boosting. *IEEE Trans. Knowledge Data Eng.* 25, 1574–1586. doi: 10.1109/TKDE.2012.89

Yang, Y., and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statist. Comput.* 25, 1129–1141. doi: 10.1007/s11222-014-9498-5

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zemanick, E. T., Sagel, S. D., and Harris, J. K. (2011). The airway microbiome in cystic fibrosis and implications for treatment. *Curr. Opin. Pediatr.* 23, 319–324. doi: 10.1097/MOP.0b013e32834604f2

Zhai, J., Hsu, C.-H., and Daye, Z. J. (2017a). Ridle for sparse regression with mandatory covariates with application to the genetic assessment of histologic grades of breast cancer. *BMC Med. Res. Methodol.* 17:12. doi: 10.1186/s12874-017-0291-y

Zhai, J., Knox, K. S., Twigg III, H. L., Zhou, H., and Zhou, J. (2017b). Exact tests of zero variance component in presence of multiple variance components with application to longitudinal microbiome study. *bioRxiv* doi: 10.1101/281246

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

Zhou, H., Alexander, D., and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statist. Comput.* 21, 261–273. doi: 10.1007/s11222-009-9166-3

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2015). MM algorithms for variance components models. *arXiv preprint arXiv:1509.07426*.

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 Microbiology

ORIGINAL RESEARCH
published: 27 June 2018
doi: 10.3389/fmicb.2018.01391

Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model

Jian Xiao [1,2], Li Chen [3], Stephen Johnson [1], Yue Yu [1], Xianyang Zhang [4] and Jun Chen [1]**

[1] Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN, United States, [2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Hubei, China, [3] Department of Health Outcomes Research and Policy, Harrison School of Pharmacy, Auburn University, Auburn, AL, United States, [4] Department of Statistics, Texas A&M University, College Station, TX, United States

Recent human microbiome studies have revealed an essential role of the human microbiome in health and disease, opening up the possibility of building microbiome-based predictive models for individualized medicine. One unique characteristic of microbiome data is the existence of a phylogenetic tree that relates all the microbial species. It has frequently been observed that a cluster or clusters of bacteria at varying phylogenetic depths are associated with some clinical or biological outcome due to shared biological function (*clustered signal*). Moreover, in many cases, we observe a community-level change, where a large number of functionally interdependent species are associated with the outcome (*dense signal*). We thus develop "glmmTree," a prediction method based on a generalized linear mixed model framework, for capturing clustered and dense microbiome signals. glmmTree uses the similarity between microbiomes, which is defined based on the microbiome composition and the phylogenetic tree, to predict the outcome. The effects of other predictive variables (e.g., age, sex) can be incorporated readily in the regression framework. Additional tuning parameters enable a data-adaptive approach to capture signals at different phylogenetic depth and abundance level. Simulation studies and real data applications demonstrated that "glmmTree" outperformed existing methods in the dense and clustered signal scenarios.

Keywords: microbiome, phylogenetic tree, kernel method, generalized mixed model, predictive model

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Bradley Stevenson,
University of Oklahoma, United States
Yijuan Hu,
Emory University, United States

****Correspondence:***
Jun Chen
chen.jun2@mayo.edu
Li Chen
li.chen@auburn.edu

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

Received: *09 February 2018*
Accepted: *06 June 2018*
Published: *27 June 2018*

Citation:
Xiao J, Chen L, Johnson S, Yu Y,
Zhang X and Chen J (2018) Predictive
Modeling of Microbiome Data Using a
Phylogeny-Regularized Generalized
Linear Mixed Model.
Front. Microbiol. 9:1391.
doi: 10.3389/fmicb.2018.01391

1. INTRODUCTION

The human microbiome, the collection of micro-organisms associated with the human body, has recently attracted substantial scientific interest due to its vital role in human health. For instance, the human gut microbiome contributes to nutrient metabolism, immune maturation and modulation, inflammatory cytokine production, and host gene regulation (Ahern et al., 2014; Schirmer et al., 2016; Pedersen et al., 2016; Fellows et al., 2018). Many diseases have been linked to dysbiosis of the microbiome ranging from metabolic disorders (e.g., obesity and type II diabetes) to autoimmune diseases (e.g., rheumatoid arthritis and multiple sclerosis) (Turnbaugh et al., 2009; Kinross et al., 2011; Cho and Blaser, 2012; Honda and Littman, 2012; Pflughoeft and Versalovic, 2012; Qin et al., 2012; Chen et al., 2016; Jangi et al., 2016). An abnormal microbiome has also been implicated in many cancer types such as colorectal, endometrial and esophageal cancers (Ahn et al., 2013; Bultman, 2014; Walther-Antonio et al., 2016; Peters et al., 2017), and a causal link has been emerging through deep mechanistic studies (Rubinstein et al., 2013; Bullman et al., 2017). In addition, the individual microbiomes may modulate drug pharmacokinetics and pharmacodynamics, contributing to drug response variations among individual patients (Haiser et al., 2014). Recently, the efficacy of cancer immune therapy has been shown to depend on the initial configuration of the gut microbiome (Gopalakrishnan et al., 2018; Matson et al., 2018; Routy et al., 2018). These findings open up the possibility of microbiome-based predictive medicine, where the microbiome data are used, potentially in conjunction with other clinic or omics data,to improve the prediction of relevant clinical outcomes.

A typical microbiome study involves collecting the microbiome samples, isolating all genomic DNA and sequencing the DNA using next-generation sequencing technologies. There are two main approaches to sequence the microbiome: gene-targeted sequencing and shotgun metagenomic sequencing (Kuczynski et al., 2011). In gene-targeted sequencing, a "fingerprint" gene that carries the taxonomic identity (e.g., 16S rRNA gene) is amplified and sequenced, while in shotgun metagenomic sequencing all genomic DNA is sequenced. Although shotgun metagenomics can profile both the taxonomic and functional content of the microbiome, the targeted approach has been more routinely employed to study the microbiome due to its lower cost and established bioinformatics pipelines. In the targeted approach, the sequencing reads are usually first clustered into operational taxonomic units (OTUs) based on the sequence similarity, via either *de novo* clustering or comparing to a reference database of OTUs (Edgar, 2013; Chen W. et al., 2013; Chen X. et al., 2018; Rideout et al., 2014). These OTUs are assumed to represent biological species at a 97% similarity level. Recently, the concept of "amplicon sequence variant" (ASV) has been proposed with the aim to cluster the sequence reads into a finer taxonomic resolution without the need for a particular similarity cutoff (e.g., 97%) (Callahan et al., 2016, 2017). After the clustering process, the sequencing reads from a targeted sequencing study are usually summarized as a count (abundance) table of the detected OTUs/ASVs. These OTUs/ASVs are all phylogenetically related, and a phylogenetic tree that reflects the evolutionary relationship can be built based on their sequence divergence (Price et al., 2010). Closely related species usually have similar biological functions, and they are likely to be associated with the outcome simultaneously, forming "clustered signals" (Martiny et al., 2015). These clustered signals can appear at a varying phylogenetic depth, resulting in clusters of different sizes (e.g., phyla and genera are at deep and shallow phylogenetic depths respectively) (Garcia et al., 2014). Thus, the phylogenetic tree provides important prior knowledge about how these species are related, which can be used to improve the efficiency of statistical analyses. Indeed, incorporation of the phylogenetic tree in the analysis has been instrumental in revealing overall community structure, identifying covariate-associated bacteria and improving the power of microbiome-wide testing (Purdom, 2011; Chen et al., 2012; Chen J. et al., 2013; Evans and Matsen, 2012; Xiao et al., 2017; Wang and Zhao, 2017).

To predict an outcome based on microbiome data, general-purpose machine learning methods, such as Random Forest and Support Vector Machine, as well as sparse regression models, such as Lasso (Tibshirani, 1996), MCP (Zhang, 1996), and Elastic Net (Zou and Trevor, 2005), have been applied (Knights et al., 2011; Statnikov et al., 2013; Pasolli et al., 2016). Although these methods are efficient in addressing the high dimensionality problem, they have a limited ability to exploit the phylogenetic structure of the microbiome data and hence may not be optimal if the signals are clustered. Many efforts have been attempted to incorporate the phylogenetic tree structure into prediction, mainly by imposing a novel phylogeny/tree-based smoothness penalty in penalized regression models. The phylogeny-based penalty encourages similar coefficients among species with respect to their phylogenetic relationship. For example, Tanaseichuk et al. (2014) used a tree-guided penalty to incorporate such structure into a penalized logistic regression framework. Chen et al. (2015) proposed a tree-based Laplacian penalty, in addition to a sparse penalty, for both classification and regression of microbiome data. These methods favor sparse and clustered signals due to their inherent sparsity assumption. However, a community-level change has frequently been observed in many physiological or pathophysiological states (Jernberg et al., 2010; Koenig et al., 2011; Milani et al., 2016), where a large number of functionally dependent species in the community are jointly associated with the outcome ("dense signal"). The "dense" signal is usually the consequence of the perturbation of the underlying microbial network, where species interact with each other to maintain a steady state (Faust and Raes, 2012). In such scenarios, although each species may have a weak effect on the outcome, the joint effects of all species may be strong. Thus, the sparsity assumption may not be desirable for "dense" microbiome signals.

In this work, we develop "glmmTree," a predictive method based on a generalized mixed model framework, for capturing clustered and dense microbiome signals. To exploit the potential phylogenetic relatedness among species, the coefficients of the species are modeled as random with the correlation structure defined based on the phylogenetic tree. Other predictive variables (e.g., age, sex) are assumed to have fixed effects. One tuning parameter in the phylogeny-induced correlation structure allows detecting signals at various phylogenetic depths, and another tuning parameter facilitates differential weighting according to the species abundances as well as capturing certain non-linear relationships. Simulation studies and real data applications demonstrate that "glmmTree" outperforms existing methods in clustered and dense-signal scenarios.

2. METHODS

2.1. A Phylogeny-Induced Correlation Structure Among OTUs

Before we develop the predictive model for microbiome data, we first introduce a phylogeny-induced correlation structure among OTUs based on an evolutionary model. We use the term "OTU" throughout to represent a basic analysis unit. Assume that we have p OTUs on a phylogenetic tree and the patristic distance between OTU (i.e., the length of the shortest path linking OTU i and j on the tree) is denoted as d_{ij}, the correlation of the traits between OTU i and j can be modeled using the following trait evolutionary model (Martins and Hansen, 1997).

$$C_{ij}(\rho) = e^{-2\rho d_{ij}}, \; i,j = 1, \ldots, p. \tag{1}$$

The parameter $\rho \in (0, \infty)$ characterizes the evolutionary rate. If $\rho = 0$, then $C_{ij} = 1$, $\forall i, j$, indicating that all the traits are the same and there is no evolution at all. If $\rho \to \infty$, then $C_{ij} \to 0$, $\forall i, j$, indicating that the evolution is so fast that there is no correlation among the OTUs. In such case, the tree is not informative. Alternatively, ρ can be interpreted as a parameter that controls the phylogenetic depth at which the OTUs are grouped: larger ρ (smaller C_{ij}) groups OTUs into clusters at a lower phylogenetic depth (a cluster is defined as a group of highly correlated OTUs). When $\rho \to \infty$, there is no grouping of the OTUs. Conceptually, the phylogenetic grouping via ρ has a similar effect as taxonomic grouping, where OTUs at different taxonomic ranks (e.g., phylum, class, order, family, genus) are grouped according to their taxonomy. Compared to taxonomic grouping, the phylogenetic grouping circumvents the difficulty of the uncertainty in taxonomy assignments and achieves far more levels of granularities by adjusting ρ.

As the square root of the phylogenetic distance d_{ij} is of Euclidean nature (de Vienne et al., 2011), $C(\rho) = (C_{ij}(\rho))_{p \times p}$ is positive definite by Bochner's theorem. In the proposed method, we recommend using $e^{-2\rho d_{ij}^2}$ to achieve an even better signal-grouping effect. Although the positive definiteness of $C(\rho)$ is no longer theoretically guaranteed, it is positive definite or close to positive definite for most applications. In case of non-positive definiteness, we can perform positive definiteness correction (Higham, 2002).

2.2. glmmTree: A Generalized Linear Mixed Model Based on a Phylogenetic Tree

We assume that there are n samples with the abundances of p OTUs being profiled. For the ith sample, let y_i denote the outcome variable of interest, which can be binary or continuous (e.g., disease status, or body mass index) , $\boldsymbol{z_i} = (z_{i1}, z_{i2}, \ldots, z_{ip})^T$ denote the normalized abundance vector of p OTUs (i.e., counts divided by the library size) for sample i, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^T$ be the $q \times 1$ vector for covariates such as gender, age and other environmental or clinical variables that have predictive values. The goal is to predict y_i by $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$.

For a continuous outcome variable, we use the linear mixed model (LMM) to build the prediction model

$$\begin{aligned} y_i &= \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + f(\boldsymbol{z}_i; \gamma)^T \boldsymbol{b} + \epsilon_i \\ \boldsymbol{b} &\sim N(0, \sigma_b^2 C(\rho)), \;\; \epsilon_i \sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{2}$$

and, for a binary outcome variable, we use the generalized linear mixed model (GLMM)

$$\begin{aligned} \text{logit}(E(y_i)) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + f(\boldsymbol{z}_i; \gamma)^T \boldsymbol{b} \\ \boldsymbol{b} \sim N(0, \sigma_b^2 C(\rho)), \end{aligned} \tag{3}$$

where β_0 is an intercept and $\boldsymbol{\beta_1} = (\beta_1, \beta_2, \ldots, \beta_q)^T$ is a $q \times 1$ vector of fixed effect regression coefficients for the q covariates, ϵ_i is the random error, $\boldsymbol{b} = (b_1, \ldots, b_p)^T$ is a $p \times 1$ vector of random effect regression coefficients, $C(\rho) = (C_{ij}(\rho))_{p \times p}$ is the phylogeny-induced correlation structure defined in the previous section, and $f(\boldsymbol{z}_i; \gamma) = (f(z_{i1}; \gamma), \ldots, f(z_{ip}; \gamma))^T$ denotes some component-wise transformation of the abundance vector with the parameter γ allowing more modeling capability.

There are two advantages assuming the OTU effects $\boldsymbol{b}$ as random. Firstly, as the sample size is typically smaller than the number of OTUs ($p > n$), treating $\boldsymbol{b}$ as fixed effects will lead to overfitting on the training data and poor generalization on the test data. To improve the generalizability of the predictive model, the regression coefficients $\boldsymbol{b}$ need to be regularized. We thus put some distributional assumption on $\boldsymbol{b}$ and assume that $\boldsymbol{b}$ comes from a multivariate normal distribution with variance-covariance structure $\sigma_b^2 C(\rho)$. The estimation procedure now switches from estimating p regression coefficients to estimating the variance component σ_b^2, which significantly reduces the number of parameters. Secondly, treating $\boldsymbol{b}$ as random effects provides the flexibility to incorporate prior structure information. For OTU data, the prior information is the phylogenetic relationship among OTUs, and closely related OTUs have a tendency to have similar effects. We incorporate such prior information using the phylogeny-induced correlation structure $C(\rho)$. It should be noted that the ratio between σ_b^2 and σ_ϵ^2 quantifies the joint (additive) OTU effects.

For the transformation function $f(\cdot)$, we propose using a power transformation, which is defined as

$$f(z_{ij}, \gamma) = \begin{cases} z_{ij}^{\gamma} & z_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where γ is an unknown constant ($\gamma \geq 0$). Similar to Box-Cox transformation (Sakia, 1992), it can potentially model a wide range of non-linear relationships between the OTU abundance and the outcome. This transformation takes into account the skewed OTU abundance distribution and allows differential weighting according to the abundance level. Smaller values of γ (e.g., 0.1) up-weight less abundant OTUs so that their effects will not be masked by those dominant OTUs when the signals are primarily in the less abundant OTU clusters. When γ approaches 0, the OTU abundance data become almost binary. In this case, only presence/absence of the OTU matters and these dominant OTUs contribute little to the outcome since they are present in most samples.

In the model, the regression coefficients β_0 and $\boldsymbol{\beta}_1$, and the variance components $\sigma_b^2, \sigma_\epsilon^2$ need to be estimated from the data. In principle, the parameters ρ and γ can also be estimated. However, in our application, we treat them as tuning parameters, and their optimal values are selected using cross-validation. We account for potential non-informativeness of the phylogenetic tree (i.e., signals are not clustered with respect to the tree) by including a very large value on the search grid of ρ.

Our phylogeny-based LMM or GLMM can be written in another form,

$$
\begin{aligned}
g(E(y_i)) &= \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + h_i \\
\boldsymbol{h} = (h_1, h_2, ..., h_n)^T &\sim MVN(0, \sigma_b^2 \mathrm{K}(\gamma, \rho))
\end{aligned} \tag{4}
$$

where $g(.)$ is the link function, $\boldsymbol{h}$ are the aggregated OTU effect (overall microbiome effect) and $\mathrm{K}(\gamma, \rho)$ is a phylogeny-based kernel matrix by evaluating the kernel function

$$K(\boldsymbol{z_i}, \boldsymbol{z_j}; \gamma, \rho) = f(\boldsymbol{z}_i; \gamma)^T \mathrm{C}(\rho) f(\boldsymbol{z}_j; \gamma)$$

at all pairs of observations. The phylogeny-based kernel function $K(\cdot, \cdot; \gamma, \rho)$ quantifies the similarity between observations in terms of OTU abundance profile ("microbiome similarity") while taking into account the phylogenetic tree structure. Similar ideas have been used to define ecological distances between microbiome samples such as the popular UniFrac distance (Lozupone and Knight, 2005). From (4), we can see that our model aims to predict the outcome based on the microbiome similarities while the tuning parameters γ, ρ are used to tailor the microbiome similarity measure to maximally reflect the outcome similarity. Since the microbiome similarity is calculated based on all OTUs, the model is expected to perform best when the signals are relatively dense, i.e., there are many outcome-associated OTUs.

Our model is closely related to the kernel machine-based semi-parametric regression model (KMR) (Liu et al., 2007, 2008)

$$g(E(y_i)) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + h_K(\boldsymbol{z_i}), \tag{5}$$

where the covariate effect is modeled parametrically, and the overall OTU effect is modeled non-parametrically through an unknown function $h_K(\cdot)$ that belongs to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ generated by the kernel function $K(\cdot, \cdot)$. It turns out that the penalized likelihood estimation for KMR is equivalent to the maximum likelihood estimation in GLMM.

2.3. Model Estimation

The parameter ρ, controlling the evolutionary rate, and the parameter γ, controlling the non-linear effect, are treated as known in model estimation. For a continuous outcome, the LMM is fitted using the restricted maximum likelihood estimation method (RMLE) as described in Kang et al. (2008). Newton-Raphson algorithm can be used to find the optimal solution. For a binary outcome, the GLMM is fitted by the penalized quasi-likelihood (PQL) method proposed by (Breslow and Clayton, 1993). PQL approximates the high-dimensional integration over $\boldsymbol{b}$ using the Laplace approximation, and the approximated likelihood function has that of a Gaussian distribution. Therefore, the PQL estimate can be obtained by fitting a series of LMMs. Details of the algorithms can be found in the Supplementary Note.

2.4. Prediction of New Observations

Once the model is fitted based on the training dataset, prediction can be made on the new observations. In this section, we describe in detail how to predict the outcome of new observations to provide more insights into our predictive model. Suppose we have n_{tr}, n_{te} observations in the training and test dataset respectively. Let $\boldsymbol{y}_{tr}, \boldsymbol{y}_{te}$ be the outcome vectors of the training and test dataset respectively, $\mathrm{X}_{tr}, \mathrm{X}_{te}$ be the design matrices for fixed effects including the intercepts and $\mathrm{Z}_{tr}, \mathrm{Z}_{te}$ be the OTU abundance matrices. We further denote $K_{tr}{=}f(\boldsymbol{Z}_{tr}; \gamma)\boldsymbol{C}(\rho)f(\boldsymbol{Z}_{tr}; \gamma)^T, K_{te}{=}f(\boldsymbol{Z}_{te}; \gamma)\boldsymbol{C}(\rho)f(\boldsymbol{Z}_{te}; \gamma)^T$ and $K_{tr,te}{=}f(\boldsymbol{Z}_{tr}; \gamma)\boldsymbol{C}(\rho)f(\boldsymbol{Z}_{te}; \gamma)^T$, which are the kernel matrices describing the microbiome similarities. We focus on the prediction of a continuous outcome and the prediction of a binary outcome can similarly be made based on the working LMM model at the convergence of the PQL algorithm.

Based on (4), the joint distribution of $\boldsymbol{y}_{tr}$ and $\boldsymbol{y}_{te}$ can be written as

$$\begin{pmatrix} \boldsymbol{y}^{tr} \\ \boldsymbol{y}^{te} \end{pmatrix} \sim \mathrm{MVN}\left\{ \begin{pmatrix} \mathrm{X}_{tr}\boldsymbol{\beta} \\ \mathrm{X}_{te}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \Sigma_{tr} & \Sigma_{tr,te} \\ \Sigma_{te,tr} & \Sigma_{te} \end{pmatrix} \right\}, \tag{6}$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, $\Sigma_{tr} = \sigma_b^2 \mathrm{K}_{tr} + \sigma_\epsilon^2 \mathrm{I}$ and $\Sigma_{te} = \sigma_b^2 \mathrm{K}_{te} + \sigma_\epsilon^2 \mathrm{I}$ are variance-covariance matrices for training and test dataset respectively, and $\Sigma_{te,tr} = \Sigma_{tr,te}^T = \sigma_b^2 \mathrm{K}_{tr,te}$ is the covariance matrix between training and test dataset. From the linear model theory, the conditional distribution of $\boldsymbol{y}_{te}$ on $\boldsymbol{y}_{tr}$ is given by

$$
\begin{aligned}
(\boldsymbol{y}_{te}|\boldsymbol{y}_{tr}) \sim \mathrm{MVN}(\mathrm{X}_{te}\boldsymbol{\beta} + \Sigma_{te,tr}\Sigma_{tr}^{-1}(\boldsymbol{y}_{tr} - \mathrm{X}_{tr}\boldsymbol{\beta}),\ \Sigma_{te} \\
- \Sigma_{te,tr}\Sigma_{tr}^{-1}\Sigma_{tr,te}).
\end{aligned} \tag{7}
$$

Thus, the prediction of $\boldsymbol{y}_{te}$ can be obtained based on

$$
\begin{aligned}
\tilde{\boldsymbol{y}}_{te} &= E[\boldsymbol{y}_{te}|\boldsymbol{y}_{tr}] \\
&= \mathrm{X}_{te}\boldsymbol{\beta} + \Sigma_{te,tr}\Sigma_{tr}^{-1}(\boldsymbol{y}_{tr} - \mathrm{X}_{tr}\boldsymbol{\beta}).
\end{aligned}
$$

Plugging in the estimates of $\boldsymbol{\beta}, \sigma_b^2$ and σ_ϵ^2 based on the training dataset, we obtain the final prediction as

$$\hat{\boldsymbol{y}}_{te} = \mathrm{X}_{te}\hat{\boldsymbol{\beta}} + \hat{\Sigma}_{te,tr}\hat{\Sigma}_{tr}^{-1}(\boldsymbol{y}_{tr} - \mathrm{X}_{tr}\hat{\boldsymbol{\beta}}).$$

Note that the prediction formula can also be written in terms of the random effects $\boldsymbol{b}$:

$$\hat{\boldsymbol{y}}_{te} = \mathrm{X}_{te}\hat{\boldsymbol{\beta}} + f(\boldsymbol{Z}_{te}; \gamma)\hat{\boldsymbol{b}},$$

where $\hat{\boldsymbol{b}}$ is the best linear unbiased predictor (BLUP), which is a smoothed estimate with respect to the phylogenetic tree (Supplementary Note).

The "glmmTree" software is available at "https://github.com/lichen-lab/glmmTree."

3. SIMULATION STUDIES

3.1. Simulation Strategy

We carried out extensive simulations to evaluate the performance of glmmTree for both continuous and binary outcomes. For the continuous outcome, we simulated 100 independent samples in the training set and 200 independent samples in the test set. For the binary outcome, we simulated 50 cases and 50 controls in the training set, and 100 cases and 100 controls in the test set. We used a Dirichlet-multinomial distribution to simulate OTU counts and generated the outcome based on the abundances of several selected OTU clusters. To objectively evaluate our predictive model, we performed a parameter sweep and investigated the effect of the cluster size (phylogenetic depth), the number of clusters (signal density) and the abundance level of the clusters on the prediction performance. The simulation studies were aimed to reveal the scenarios under which our model performed favorably and also identify potential "blind spots" of our model.

3.1.1. Simulating OTU Abundance Data

We generated the OTU counts using a Dirichlet-multinomial distribution with the parameters (the mean proportion vector and the dispersion parameter ϕ) estimated based on a real OTU dataset from a study of the microbiome of the human upper respiratory tract (Charlson et al., 2010; Chen and Li, 2013), which contains the counts of 778 OTUs from 60 samples, together with a phylogenetic tree describing the evolutionary relationship among the 778 OTUs. For each sample, the total read count was drawn from a negative binomial distribution with mean 5000 and dispersion 25. The OTU counts were normalized into OTU proportions (z) by dividing the total read counts.

3.1.2. Constructing Outcome-Associated OTU Clusters

The underlying relationship between the outcome and the microbiome is complex. The outcome-associated OTUs ("aOTUs") can be clustered at different phylogenetic depths (deep or shallow), creating OTU clusters ("aClusters") of different sizes. It is also possible that the aOTUs are simply not phylogenetically related. In such case, each aOTU constitutes an aCluster of size 1. The signal density (number of aClusters) can also vary depending on the outcome. Finally, aClusters can be abundant or rare since both rare and abundant taxa have been observed to associate with the outcome. We thus studied the effects of all these parameters in the simulation.

To construct aClusters with a different level of cluster size, signal density and abundance, 778 OTUs were first grouped into m clusters based on their patristic distances on the phylogenetic tree.

We assumed that there were m_c ($m \times s\%$) aClusters and $s\%$ represents the signal density. For given m and m_c, we chose aClusters of different abundance level (a). The simulation strategy is illustrated in **Figure 1** and the detailed settings for cluster size, signal density and abundance are presented below:

- **Cluster size ($\boldsymbol{m}$)**:
  The 778 OTUs were partitioned into m clusters using the partitioning-around-medoids (PAM) algorithm based on the patristic distances among OTUs (Chen et al., 2012). We considered $m \in (10, 100, 778)$, representing large, medium and small OTU clusters, and aClusters were selected from these OTU clusters. Note that when m=778, the aOTUs are not phylogenetically related and the phylogenetic tree is not informative for prediction.
- **Signal density ($\boldsymbol{s}\%$)**: We selected $s\% \in (10\%, 20\%, 40\%)$ for m=10, $s\% \in (1\%, 5\%, 25\%)$ for m=100 and $s\% \in (1\%, 5\%, 30\%)$ for m=778 to represent low, medium and high signal density respectively. The number of aClusters m_c was taken to be the integer part of $m \times s\%$.
- **Abundance ($\boldsymbol{a}$)**: Given m and m_c, we had $\binom{m}{m_c}$ choices of aClusters. To obtain low, medium and high abundance level, we randomly picked m_c clusters from m clusters 1000 times and recorded their cumulative abundances a_t ($t = 1, \cdots, 1000$). We chose m_c aClusters of high, medium and low abundance with abundance $\max(a_t)$, $\text{median}(\text{a}_\text{t})$, $\min(\text{a}_\text{t})$, $\text{t} = 1, ..., 1000$, respectively.

3.1.3. Generating the Outcome Based on the Abundance of AClusters

Denote C_l as the set containing the indices of the lth aCluster, $l \in \{1, \cdots, m_c\}$, and η_i be the expected outcome value for sample i. We first generated η_i based on the following linear relationship

$$\eta_i = \beta_0 + \sum_{l=1}^{m_c}(\sum_{k \in C_l} z_{ik})b_l \qquad b_l \sim N(0, \sigma_b^2) \tag{8}$$

For a continuous outcome,

$$y_i = \eta_i + \epsilon_i, \ \ \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{9}$$

For a binary outcome,

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad y_i \sim Bernoulli(\pi_i) \tag{10}$$

Note that we assigned the same coefficient for OTUs within the same cluster to create clustered signals. The variance σ_b^2 can be adjusted to control the signal-to-noise ratio. Without loss of generality, σ_b^2 was set to be 2 for the continuous outcome and 4 for the binary outcome. The error variance σ_ϵ^2 for the continuous outcome was chosen to be $\frac{1}{4}\text{var}(\boldsymbol{Zb})$ so that the OTUs jointly explain 80% of the outcome variability.

To study the prediction performance under potential non-linearity, we also simulated non-linear relationships, where we use $f(z_{ik})$ instead of z_{ik} to generate the outcome. We specifically investigated when $f(z_{ik}) = z_{ik}^{0.5}$, which attenuates the effect of highly abundant OTUs, and $f(z_{ik}) = 1(\text{if} \ \ z_{ik} \neq 0)$, which represents the scenario where only the presence/absence of the OTU affects the outcome.

3.2. Competing Methods, Model Selection and Evaluation

3.2.1. Competing Methods

We compared glmmTree to Lasso, MCP and Elastic Net (Enet), three sparse regression models with no consideration of the phylogenetic structure. Particularly, Elastic Net encourages the data-driven smoothing via L_2 penalty. We also compared glmmTree to a phylogeny-constrained sparse regression

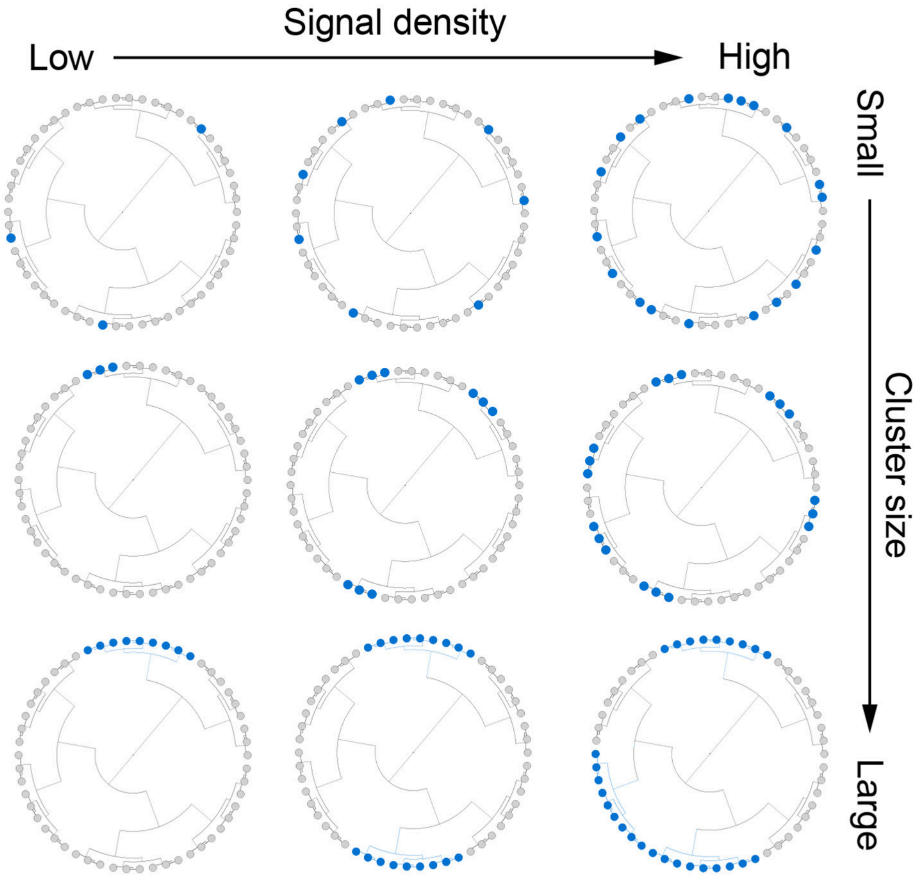


FIGURE 1 | Illustration for the simulation strategy. We simulate outcome-associated OTU clusters (aClusters) of different cluster size (**top** to **bottom**) and signal density (**left** to **right**). We also vary the abundance level of the aClusters (not shown).

model (Chen et al., 2015) as a representative of tree-structure penalized regression models. The method uses the same phylogeny-induced correlation structure as in glmmTree but encourages the phylogeny-driven smoothing based on the inverse correlation matrix instead of the usual Laplacian matrix. We thus termed it Sparse Inverse Correlation Shrinkage method (SICS). Besides those sparse regression models, we also compared glmmTree to Random Forest (RF), which has been demonstrated a superior prediction performance in various microbiome datasets. Finally, we compared to a regular kernel-based GLMM (glmmTree.Reg) to evaluate the benefit of exploiting the phylogenetic tree in prediction.

3.2.2. Model Selection and Evaluation

For glmmTree, the tuning parameters (γ, ρ) are used to control the phylogenetic depth and non-linear effect and need to be tuned. We searched ρ on the grid $\underbrace{\{0, 2^{-5}, 2^{-4}, 2^{-3}, \cdots, 2^4, 2^5\}}_{11}$ while γ was tuned on the grid $\underbrace{\{0, 0.01, 0.1, 0.3, 0.5, 0.7, ..., 1.9\}}_{12}$. glmmTree.Reg was achieved by fixing ρ at a very large value (10^4).

Box 1 | Tuning parameter settings in different methods.

- Lasso: *glmnet* R package, all parameters were set as the default.
- Elastic Net: *glmnet* R package, all parameters were set as the default.
- MCP: *ncvreg* R package, all parameters were set as the default
- SICS: *glmgraph* R package, the search grid for ρ was the same as glmmTree, the tuning parameter for the smoothness penalty was selected from $\underbrace{\{0, 2^{-5}, 2^{-4}, 2^{-3}, \cdots, 2^4, 2^5\}}_{11}$, other parameters were set as default.
- Random Forest: *randomForest* R package, parameters were set as default.

The details of specific software packages used and their parameter settings for competing methods are shown in **Box 1**.

Tuning parameter selection was based on five-fold cross-validation (CV), where the training samples were randomly divided into five folds with four folds used for model fitting and the remaining one for calculating some CV criterion. We used PMSE (Predicted Mean Square Error) as the CV criterion for a continuous outcome and AUC (Area Under the Curve) for a binary outcome. Once the optimal values of the tuning parameters were selected, we fit the model using all training sample (n=100) and then evaluated the prediction performance on the test dataset (n=200). Although we used PMSE and AUC for tuning parameter selection, we focused on R^2, which quantifies the correlation between the predicted outcome and the observed outcome and ranges from 0 (no correlation) to 1 (perfect correlation), to evaluate the prediction performance. Specifically, for a continuous outcome, R^2 is defined as

$$\mathrm{R}^2 = \frac{\{\sum_{i=1}^{n_{te}}(\hat{y}_{te,i} - \bar{\hat{y}}_{te})(y_{te,i} - \bar{y}_{te})\}^2}{\sum_{i=1}^{n_{te}}(\hat{y}_{te,i} - \bar{\hat{y}}_{te,})^2 \sum_{i=1}^{n_{te}}(y_{te,i} - \bar{y}_{te})^2},$$

where $\bar{\hat{y}}, \bar{y}$ are the sample means. For the binary-version R^2, we substitute $\hat{y}_{te,i}$ with the predicted probability $\hat{P}_{te,i}$. Each simulation was repeated 50 times and means and standard errors were reported.

3.3. Simulation Results

3.3.1. Results for the Continuous Outcome.

We first evaluated the performance of different methods across different cluster sizes and signal densities when the abundance of the aClusters was high (**Figure 2**). We observed a general decrease in performance for all methods when the signal density increased. This trend is explained by a result of decreasing individual effects as we increased the number of aOTUs since we fixed the percentage of variability explained by OTUs (80%) across parameter settings. The reduction in individual effects was unfavorable for all methods. When the aCluster was large, i.e., the signals were highly clustered, glmmTree outperformed other methods substantially. Particularly, glmmTree had a clear advantage over glmmTree.Reg, which did not account for the phylogenetic structure, indicating the benefit of using phylogenetic information to improve prediction. It was also significantly better than the sparse regression methods and RF across different levels of signal density. The unfavorable performance of these sparse regression methods was due to the weak individual effects of these aOTUs in the large cluster. In such "many OTUs, weak effects" scenario, sparse regression methods tended to have a low sensitivity and specificity to identify these aOTUs, which led to poor prediction performance. As the cluster size decreased, the phylogenetic signal became weaker, and the difference of performance between glmmTree and other methods diminished accordingly. However, glmmTree still performed better than sparse regression methods when the signal was dense. This was due to the fact that glmmTree did not assume sparsity in the model, and when the signal became dense, the irrelevant OTUs did not seriously corrupt the overall microbiome similarity, upon which the glmmTree was based. It should be noted that glmmTree and glmmTree.Reg had performance similar to those sparse regression methods in their most unfavorable setting, where a small number of phylogenetically non-related OTUs were associated with the outcome (**Figure 2A**, upper left). The comparable performance is explained by the high abundance of the aOTUs, which dominated those rare and less abundant OTUs in determining the microbiome similarity.

As we decreased the abundance of the aClusters to be "medium" (**Figure 2B**), glmmTree still excelled in highly clustered signals across different signal densities, but its prediction performance deteriorated significantly as the signal density became lower and the size of aCluster became smaller. When the signals were not phylogenetically related (**Figure 2B**, top row), sparse regression models and RF performed better than glmmTree. As these phylogenetically non-related signals grew more sparse, glmmTree had very low predictive power. A similar trend was observed when the abundance of aClusters was "low" (Figure S1). In this scenario, the phylogeny-regularized sparse regression method (SICS) outperformed the other sparse regression methods. In summary, no methods dominates in all settings and glmmTree has a performance edge over other competing methods when the signal is *dense*, *clustered* and/or *abundant*.

In glmmTree, we included two tuning parameters γ, which up-weights or down-weights the effect of abundant OTUs, and ρ, which controls the phylogenetic depth of the signal. These two tuning parameters are used to exploit various signal structures for microbiome data. It is interesting to observe the patterns of the selected values across simulation settings. We plotted the distribution of selected γ and ρ values over the fifty simulation runs across different levels of cluster size, signal density and abundance for the continuous outcome (**Figure 3**). As expected, smaller values of γ tended to be selected for "low-abundance" scenarios, where the outcome was associated with less abundant aClusters. Smaller γ values up-weighted the effects of less abundant OTUs and hence amplified their weak signals (**Figure 3A**). γ had the stronger impact when the phylogenetic signal was weak (i.e., the OTUs were less phylogenetically related). On the other hand, smaller ρ values were selected for larger clusters, where the signals were at a deeper phylogenetic depth (**Figure 3B**). Therefore, the inclusion of these two tuning parameters improved the model flexibility.

To study the robustness of glmmTree to tree mis-specification, we generated "noisy" trees by randomly permuting different percentages of the rows/columns of the tree-induced distance matrices. As we increased the percentage from 25 to 75%, the performance of glmmTree decreased accordingly, but it was still more powerful than glmmTree.Reg, which did not use tree information (Figure S2). As the tuning parameter ρ approaches infinity, glmmTree is reduced to glmmTree.Reg. Therefore, the performance of glmmTree is expected to be close to glmmTree.Reg when the tree is severely mis-specified. We next studied the performance of glmmTree under much lower percentages of variability explained by OTUs (50% and 33%). As we lowered the signal-noise-ratio (SNR), the performance of all methods deteriorate but the same trend has been observed as in the high SNR scenario (Figure S3).

3.3.2. Results for the Binary Outcome.

We repeated the same simulations for the binary outcome and present the results in **Figure 4** and Figure S4. Compared to the continuous outcome-based simulations, the performance for all methods deteriorated faster when the aClusters became less abundant and more sparse. Nevertheless, a similar trend persisted: glmmTree had the best performance under clustered

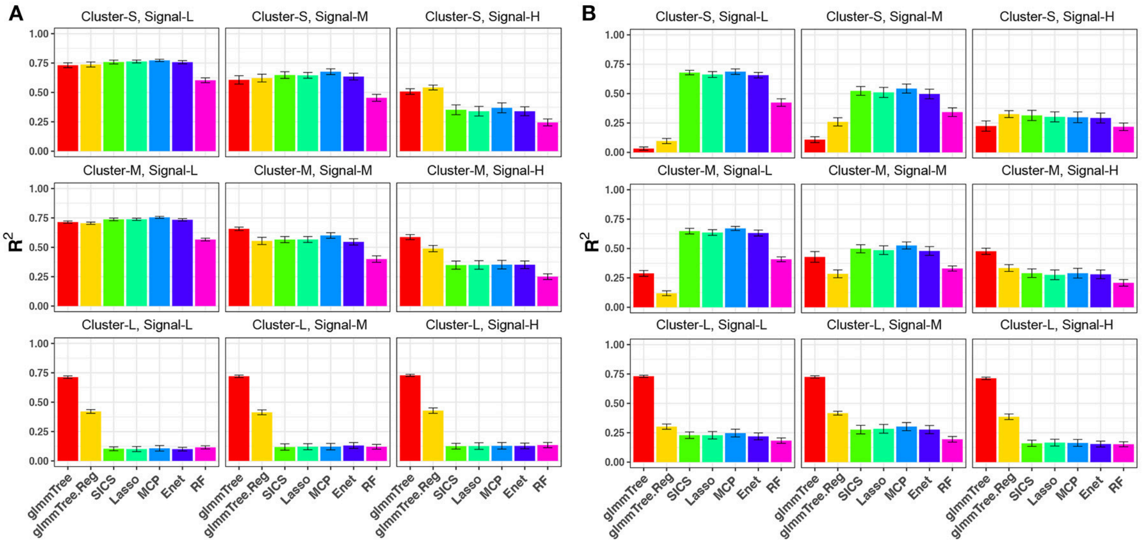


FIGURE 2 | R^2 for continuous-outcome simulations across different levels of cluster size and signal density. The abundance of associated OTU clusters is chosen to be high **(A)** and medium **(B)**. Cluster-S, -M, and -L represent small, medium and large clusters, and Signal-L, -M, and -H represent low, medium and high signal density, respectively.

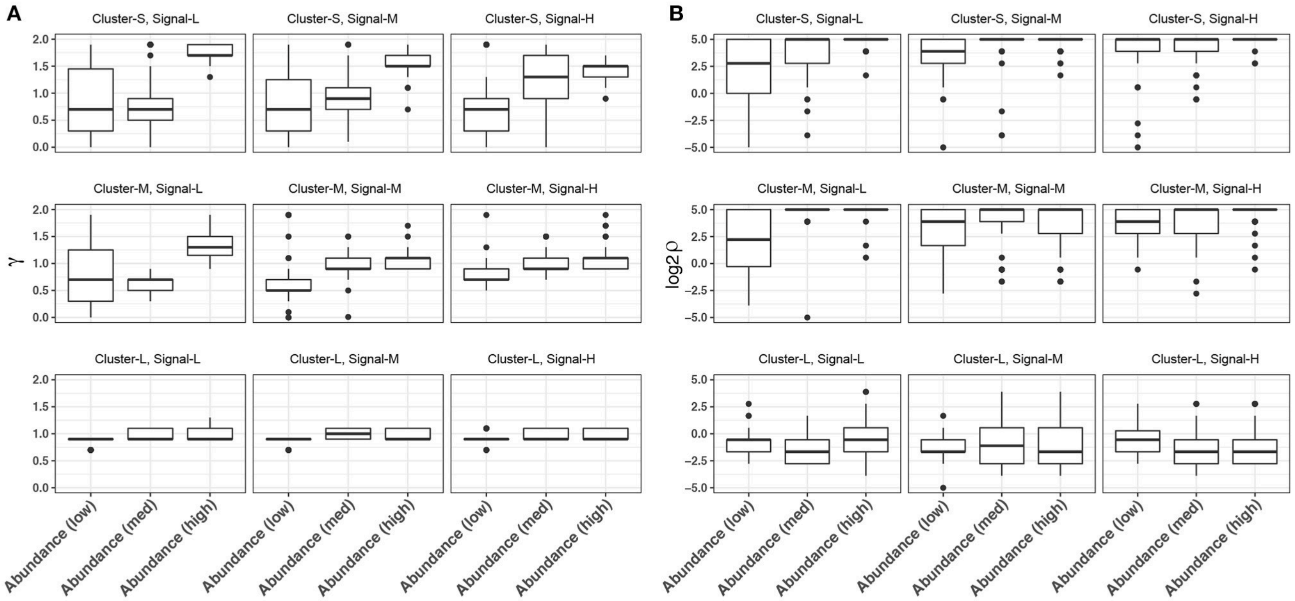


FIGURE 3 | Distribution of the selected tuning parameter γ **(A)** and ρ **(B)** across different levels of cluster size, signal density and abundance for continuous-outcome simulations. Cluster-S, -M, and -L represent small, medium and large clusters, and Signal-L, -M, and -H represent low, medium and high signal density, respectively.

and dense signals, and abundant aClusters further improved its performance.

3.3.3. Accommodation for Non-linear Signals

The conclusions in the previous simulations were based on linear signals. Since the relationship between the microbiome and the outcome is very complex, traditional linear models may fail to capture non-linear microbiome effects. Besides the differential weighting function, the tuning parameter γ can also accommodate a wide range of non-linear effects. To illustrate this point, we performed additional simulations based on non-linear signals and compared the prediction performance to glmmTree with a fixed gamma value (γ=1). Specifically, we investigated two types of non-linear relationships, in which the outcome was generated based on (1) the OTU presence/absence and (2) square-root transformed OTU abundances, respectively.

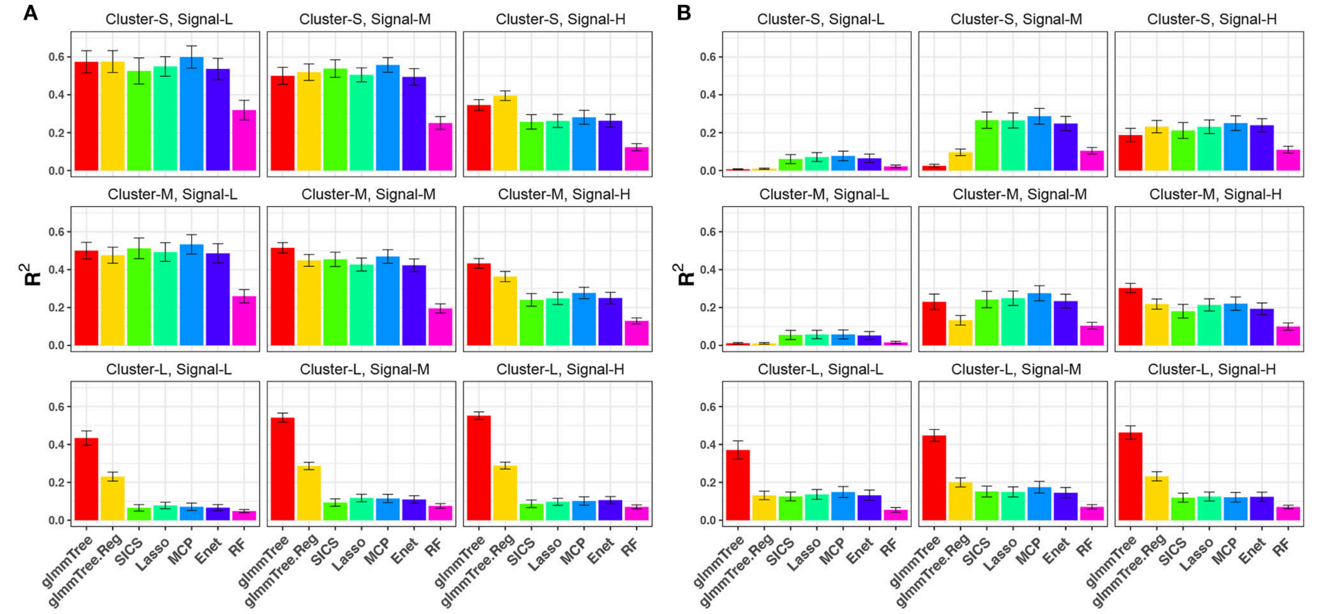


FIGURE 4 | R^2 for binary-outcome simulations across different levels of cluster size and signal density. The abundance of associated OTU clusters is chosen to be high **(A)** and medium **(B)**. Cluster-S, -M, and -L represent small, medium and large clusters, and Signal-L, -M, and -H represent low, medium and high signal density, respectively.

Without loss of generality, we set the scenario to be high abundance, large cluster and low signal density. The simulation results are presented in **Figure 5**. Clearly, glmmTree achieved a significantly higher R^2 than glmmTree without γ tuning in both non-linear scenarios for both continuous and binary outcomes. When the outcome depended on the OTU presence/absence, glmmTree without γ tuning was powerless: the R^2 was close to 0. In contrast, glmmTree with γ tuning performed substantially better since γ was usually tuned to be close to 0 to accommodate such non-linearity. When the outcome depended on the square-root transformed OTU abundances, glmmTree without γ tuning achieved some predictive power, but was still much less powerful than glmmTree with γ tuning. Therefore, glmmTree can also capture non-linear signals with the imbedded power transformation.

4. APPLICATION OF GLMMTREE TO PREDICTING CHRONOLOGICAL AGE BASED ON THE HUMAN GUT MICROBIOME

We applied glmmTree to a study investigating how the gut microbiome differs across age and geography (Yatsunenko et al., 2012). The study consisted of 531 individuals, among which 115 individuals were from Malawi, 100 individuals were from Venezuela, and 316 individuals were from the USA. The gut microbiota of these individuals was profiled using 16S rRNA gene targeted sequencing. The dataset was available for download from Qiita (https://qiita.ucsd.edu/) with study ID 850, where the sequence data was processed by the QIIME pipeline (reference-based approach). A total of 14,170 OTUs were produced for this dataset. To demonstrate the performance of glmmTree, we used the 316 individuals from the USA for age prediction.

The complexity of the real data required us to properly normalize, transform and filter the data before applying various predictive tools. Let $(c_{ij})_{p\times n}$ be the observed count matrix. We carried out a series of pre-processing steps before applying various prediction methods:

1. Sample filtering to remove outlier samples. We calculated the Bray-Curtis distance between samples. Denote d_{jk} the distance between sample j and k. For each sample j, we calculated the median distance from sample j to other samples, denoted as $m_j = Median_{k\neq j}(d_{jk})$. An outlier index o_j for sample j was defined as $o_j = m_j/Median_k(m_k)$. We removed samples with $o_j > 2$ (8 samples removed).
2. OTU filtering to remove less informative and noisy OTUs and reduce dimensionality. We imposed two filters: (1) OTU prevalence < 10%, and (2) Median non-zero counts < 10.
3. Normalization to address variable library sizes. We used GMPR normalization, which is developed specifically for zero-inflated count data (Chen L. et al., 2018). For each sample, we calculated a GMPR size factor s_j and the normalized counts were then divided by s_j. The normalized counts are denoted as $(\tilde{c}_{ij})_{p\times n}$.
4. Winsorization to replace outlier counts. For each taxon i, we calculated the 97% quantile $q_i^{0.97}$ based on $\tilde{c}_{ij}(j=1\cdots n)$, and replaced $\tilde{c}_{ij} > q_i^{0.97}$ with $q_i^{0.97}$. This procedure has shown to be effective in reducing false positives in the context of differential abundance analysis (Chen J. et al., 2018).

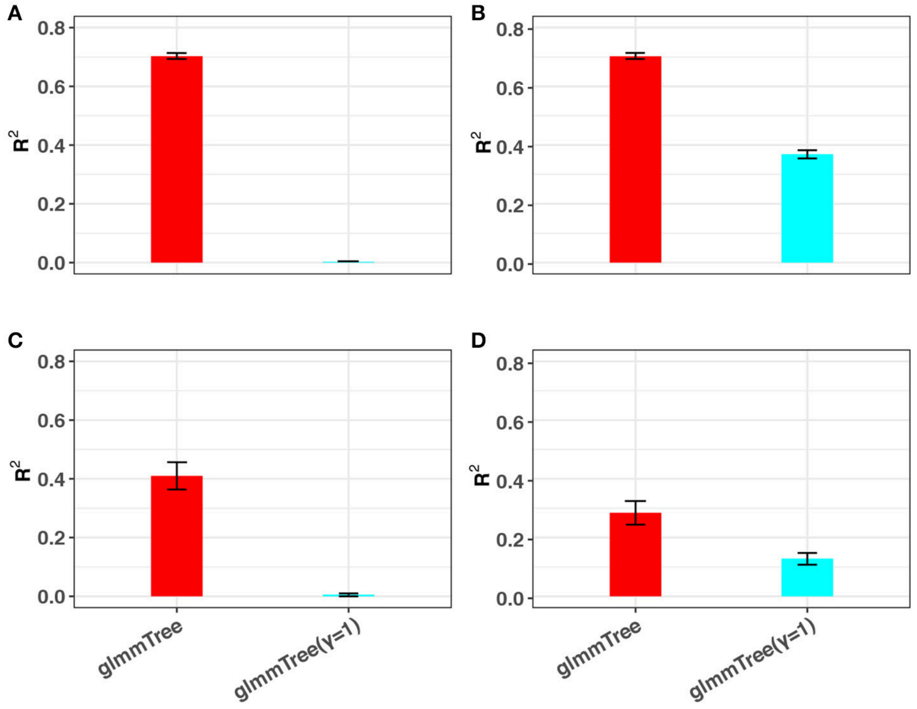


FIGURE 5 | The ability of glmmTree to capture non-linear effects through the tuning parameter γ. glmmTree with tunable γ (red) is compared to glmmTree with fixed $\gamma = 1$ (blue). R^2 is used to evaluate the performance for continuous **(A,B)** and binary **(C,D)** outcomes when the outcome is generated based on OTU presence/absence **(A,C)** and square-root transformed OTU abundances **(B,D)**.

5. Transformation to reduce the influence of highly abundant taxa counts. We used the commonly used square-root transformation.
6. We further used square-root transformation on the continuous age variable to better capture the underlying relationship.

These proprocessing steps were used to make the microbiome data more amenable to predictive modeling, and could improve the performance of sparse regression methods such as Lasso (Figure S5). After the processing steps, we were left with 308 individuals and 1087 OTUs. We first evaluated the prediction performance by treating age as a continuous outcome. To demonstrate the performance with binary outcomes, we classified the individuals into three age groups: baby (age $\leq$ 3 years, $n = 54$), child ($3 <$ age < 18 years, $n = 125$) and adult (age $\geq$ 18 years, $n = 129$), and evaluated the prediction performance based on the baby and child age group. The guidance of the group division and choice was based on the observation that the microbiome change begins to slow down after three years old, and the child microbiome is more similar to the adult microbiome (Yatsunenko et al., 2012). We included the prediction of baby vs. child in the main text and the prediction of child vs. adult in the Supplementary File.

We compared glmmTree to SICS, Lasso, MCP, Elastic Net and Random Forest. Tuning parameter selection was based on cross-validation (CV) as in the simulation.

To have an objective evaluation of the prediction performance, we randomly divided the dataset fifty times into five folds: four folds were used for training (with nested CV) and the remaining one fold for testing. R^2 and PMSE were used as metrics for the continuous outcome, while R^2 and AUC were used for the binary outcome. The results are presented in **Figure 6**. glmmTree achieved the best performance for continuous age prediction as indicated by the highest R^2 and lowest PMSE, followed by SICS and Elastic Net. For baby vs. child prediction, glmmTree still achieved the highest R^2 and AUC, followed by Elastic Net and Random Forest. For child vs. adult prediction, glmmTree and Elastic net achieved the best performance (Figure S6). To verify if the improvement of prediction was significant, we performed paired Wilcoxon signed-rank tests between glmmTree and other methods based on R^2, PMSE and AUC obtained from the fifty random divisions. For continuous age prediction, glmmTree achieved significantly higher R^2, and significantly lower PMSE than other methods (P-value < 0.05). For baby vs. child prediction, glmmTree achieved significantly higher AUC than other methods, and significantly higher R^2 than other methods except Elastic Net. For child vs. adult prediction, glmmTree achieved significantly higher AUC and R^2 than other methods except Elastic net. Overall, glmmTree performed the best for both the continuous and binary age outcome on this dataset.

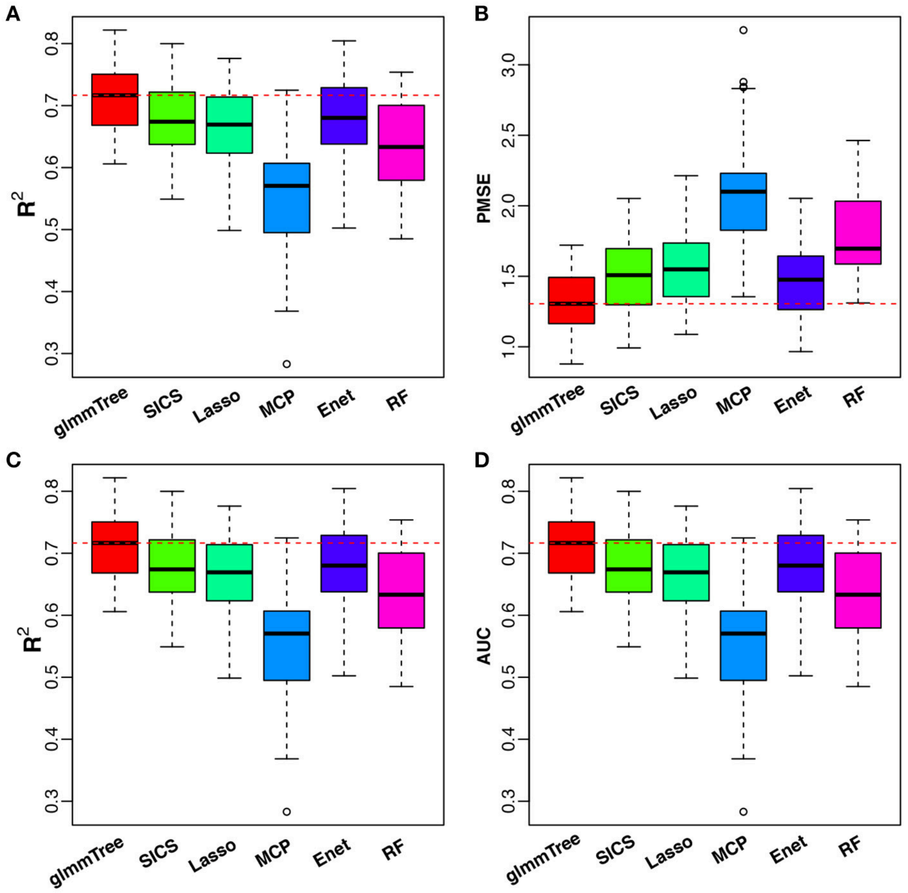


FIGURE 6 | Performance comparison for age prediction. All USA samples are used in continuous age prediction **(A,B)**. Binary prediction is based on the two age groups: baby (0 to 3 years old) and child (3 to 18 years old) **(C,D)**. Red dashed line indicates the median value of various performance measures for glmmTree.

5. DISCUSSION

One of the challenges for predictive modeling of microbiome data is the utilization of the phylogenetic tree. As microbiome profiling experiments produce increasingly higher taxonomic resolutions such as strain-level resolution (Truong et al., 2015; Callahan et al., 2016), incorporating the phylogenetic tree information becomes even more important. The phylogenetic tree provides a principled way to pool signals and directs the analysis to the most relevant parameter space, which is essential to counter the "curse of dimensionality." Previous work indicates that predictive models could benefit from the incorporation of the phylogenetic tree through the use of tree-induced smoothness penalty (Tanaseichuk et al., 2014; Chen et al., 2015; Wang and Zhao, 2017). These models usually induce a sparse solution and are hence efficient to detect sparse and clustered signals. In this work, we propose to utilize the phylogenetic tree to detect dense and clustered signals. This is achieved by assuming the OTU effects as random in a GLMM framework, and that the OTU random effects follow a multivariate normal distribution with the correlation structure defined based on the phylogenetic tree.

We performed comprehensive simulations to investigate the performance of the proposed method at varying cluster sizes, signal densities and taxa abundances. Simulation studies demonstrated that glmmTree favors dense and clustered signals or signals from abundant OTUs, compared to sparse regression models, which has a competitive performance for sparse signals, particularly from those less abundant OTUs. By using a power transformation, glmmTree can capture a wide range of non-linear effects including the biologically relevant scenario where the outcome depends on the presence/absence of the OTUs. Human microbiome studies have frequently found that the species richness (α-diversity) were associated with some phenotypic traits (Le Chatelier et al., 2013). Therefore, capturing the signals on the presence/absence level should not be overlooked.

Our work is closely related to the recently proposed kernel penalized regression framework (Randolph et al., 2015), which provides a theoretic framework to incorporate a variety of extrinsic information, such as phylogeny, into penalized regression models. For microbiome data applications, Randolph et al. (2015) illustrated their method using a kernel-based on UniFrac distances. In our work, we took a further step and optimized the microbiome-based kernel to be capable of capturing clustered signals at various phylogenetic depth as well as accommodating non-linearity. Moreover, our model is based on the generalized linear model, which can handle non-Gaussian outcomes while adjusting for covariates easily.

As the microbiome field matures, more complex study designs such as family and longitudinal studies have been used to study the human microbiome in relation to various clinical and biological variables. These studies are efficient to control potential confounders such as genetics and diet, and are also more powerful than studies based on independent sampling. Although our framework is developed mainly for independent data, it could be modified to accommodate such clustered data by incorporating additional cluster-level random effects. Similar algorithms (i.e., PQL) could be used to fit these multiple random effects model.

The effectiveness of the proposed method depends on the reliability of the phylogenetic tree, which can be very noisy or non-informative. Although our method is robust to tree mis-specification via the tuning parameter ρ, its performance will not be optimal if the tree is severely mis-specified. In this case, other types of kernels without using the tree, such as the radial basis function (RBF) kernel (Shawe-Taylor and Cristianini, 2004), may be more powerful. A composite kernel that combines the tree-based and non-tree-based kernels may increase the robustness of our method for detecting various kinds of dense signals. Furthermore, since the underlying signal structure is unknown for real applications, an ensemble approach incorporating representative prediction methods targeted to different signal structures (e.g., dense vs. sparse) is more likely to provide an even more robust prediction. We leave these extensions as our future work.

AUTHOR CONTRIBUTIONS

JX analyzed the data, drafted the paper, prepared figures and tables, reviewed drafts of the paper. LC analyzed the data, drafted the paper, prepared figures and tables, wrote the software, reviewed drafts of the paper. SJ revised drafts of the paper. YY contributed to the revision of the paper. XZ contributed substantial expertise to improve the paper and revised the paper. JC conceived and designed the experiments, analyzed the data, wrote the paper, wrote the software, prepared figures and tables.

ACKNOWLEDGMENTS

This work was supported by Mayo Clinic Gerstner Family Career Development Awards, Mayo Clinic Center for Individualized Medicine, U01 FD005875, Food and Drug Administration and the National Natural Science Foundation of China (no.61773401).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01391/full#supplementary-material

REFERENCES

Ahern, P. P., Faith, J. J., and Gordon, J. I. (2014). Mining the human gut microbiota for effector strains that shape the immune system. *Immunity* 40, 815–823. doi: 10.1016/j.immuni.2014.05.012

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., et al. (2013). Human gut microbiome and risk for colorectal cancer. *J. Natl. Cancer Inst.* 105, 1907–1911. doi: 10.1093/jnci/djt300

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

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of fusobacterium persistence and antibiotic response in colorectal cancer. *Science* 358, 1443–1448. doi: 10.1126/science.aal5240

Bultman, S. J. (2014). Emerging roles of the microbiome in cancer. *Carcinogenesis* 35, 249–255. doi: 10.1093/carcin/bgt392

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., and Holmes, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE* 5:e15216. doi: 10.1371/journal.pone.0015216

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

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

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., et al. (2018). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* 34, 643–651. doi: 10.1093/bioinformatics/btx650

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

Chen, J., Wright, K., Davis, J. M., Jeraldo, P., Marietta, E. V., Murray, J., et al. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* 8, 43. doi: 10.1186/s13073-016-0299-7

Chen, L., Liu, H., Kocher, J. P., Li, H., and Chen, J. (2015). glmgraph: an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* 31, 3991–3993. doi: 10.1093/bioinformatics/btv497

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

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013). A comparison of methods for clustering 16s rrna sequences into otus. *PLoS ONE* 8:e70837. doi: 10.1371/journal.pone.0070837

Chen, X., Johnson, S., Jeraldo, P., Wang, J., Chia, N., Kocher, J. A., et al. (2018). Hybrid-denovo: a *de novo* otu-picking pipeline integrating single-end and paired-end 16s sequence tags. *Gigascience* 7, 1–7. doi: 10.1093/gigascience/gix129

Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* 13, 260–270. doi: 10.1038/nrg3182

de Vienne, D., Aguileta, G., and Ollier, S. (2011). Euclidean nature of phylogenetic distance matrices. *Syst. Biol.* 60, 826–832. doi: 10.1093/sysbio/syr066

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

Evans, S. N., and Matsen, F. A. (2012). The phylogenetic kantorovich-rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74, 569–592. doi: 10.1111/j.1467-9868.2011.01018.x

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

Fellows, R., Denizot, J., Stellato, C., Cuomo, A., Jain, P., Stoyanova, E., et al. (2018). Microbiota derived short chain fatty acids promote histone crotonylation in the colon through histone deacetylases. *Nat. Commun.* 9, 105. doi: 10.1038/s41467-017-02651-5

Garcia, T. P., Muller, S., Carroll, R. J., and Walzem, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* 30, 831–837. doi: 10.1093/bioinformatics/btt608

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., et al. (2018). Gut microbiome modulates response to anti-pd-1 immunotherapy in melanoma patients. *Science* 359, 97–103. doi: 10.1126/science.aan4236

Haiser, H. J., Seim, K. L., Balskus, E. P., and Turnbaugh, P. J. (2014). Mechanistic insight into digoxin inactivation by eggerthella lenta augments our understanding of its pharmacokinetics. *Gut. Microbes* 5, 233–238. doi: 10.4161/gmic.27915

Higham, N. (2002). Computing the nearest correlation matrixa problem from finance. *IMA J. Numer. Anal.* 22, 329–343. doi: 10.1093/imanum/22.3.329

Honda, K., and Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annu. Rev. Immunol.* 30, 759–795. doi: 10.1146/annurev-immunol-020711-074937

Jangi, S., Gandhi, R., Cox, L. M., Li, N., von Glehn, F., Yan, R., et al. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* 7:12015. doi: 10.1038/ncomms12015

Jernberg, C., Lofmark, S., Edlund, C., and Jansson, J. K. (2010). Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* 156(Pt 11), 3216–3223. doi: 10.1099/mic.0.040618-0

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

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

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

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108 (Suppl. 1), 4578–4585. doi: 10.1073/pnas.1000081107

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58. doi: 10.1038/nrg3129

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

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* 9:292. doi: 10.1186/1471-2105-9-292

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

Lozupone, C., and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Martins, E. P., and Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* 149, 646–667. doi: 10.1086/286013

Martiny, J. B., Jones, S. E., Lennon, J. T., and Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. doi: 10.1126/science.aac9323

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M. L., et al. (2018). The commensal microbiome is associated with anti-pd-1 efficacy in metastatic melanoma patients. *Science* 359, 104–108. doi: 10.1126/science.aao3290

Milani, C., Ticinesi, A., Gerritsen, J., Nouvenne, A., Lugli, G. A., Mancabelli, L., et al. (2016). Gut microbiota composition and clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. *Sci. Rep.* 6:25945. doi: 10.1038/srep25945

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

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535, 376–381. doi: 10.1038/nature18646

Peters, B. A., Wu, J., Pei, Z., Yang, L., Purdue, M. P., Freedman, N. D., et al. (2017). Oral microbiome composition reflects prospective risk for esophageal cancers. *Cancer Res.* 77, 6777–6787. doi: 10.1158/0008-5472.CAN-17-1296

Pflughoeft, K. J., and Versalovic, J. (2012). Human microbiome in health and disease. *Annu. Rev. Pathol.* 7, 99–122. doi: 10.1146/annurev-pathol-011811-132421

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Purdom, E. (2011). Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* 5, 2326–2358. doi: 10.1214/10-AOAS402

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

Randolph, T., Zhao, S., Copeland, W., Hullar, M., and Shojaie, A. (2015). Kernel-penalized regression for analysis of microbiome data. arXiv preprint arXiv:1511.00297.

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., et al. (2014). Subsampled open-reference clustering creates consistent, comprehensive otu definitions and scales to billions of sequences. *PeerJ* 2:e545. doi: 10.7717/peerj.545

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillere, R., et al. (2018). Gut microbiome influences efficacy of pd-1-based immunotherapy against epithelial tumors. *Science* 359, 91–97. doi: 10.1126/science.aan3706

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating e-cadherin/beta-catenin signaling via its fada adhesin. *Cell Host Microbe* 14, 195–206. doi: 10.1016/j.chom.2013.07.012

Sakia, R. (1992). The box-cox transformation technique: a review. *Statistician* 63, 169–178. doi: 10.2307/2348250

Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., et al. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* 167, 1125–1136. doi: 10.1016/j.cell.2016.10.020

Shawe-Taylor, J., and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.

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

Tanaseichuk, O., Borneman, J., and Jiang, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* 30, 449–456. doi: 10.1093/bioinformatics/btt700

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

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods.* 12, 902–903. doi: 10.1038/nmeth.3589

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

Walther-Antonio, M. R., Chen, J., Multinu, F., Hokenstad, A., Distad, T. J., Cheek, E. H., et al. (2016). Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Med.* 8, 122. doi: 10.1186/s13073-016-0368-y

Wang, T., and Zhao, H. (2017). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Am. Stat. Assoc.* 112, 1022–1031. doi: 10.1080/01621459.2016.1270213

Xiao, J., Cao, H., and Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* 33, 2873–2881. doi: 10.1093/bioinformatics/btx311

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

Zhang, C. H. (1996). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 58, 267–288.

Zou, H., and Trevor, H. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 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.

frontiers in Microbiology

METHODS
published: 03 May 2018
doi: 10.3389/fmicb.2018.00872

Identifying *Group-Specific* Sequences for Microbial Communities Using Long *k*-mer Sequence Signatures

Ying Wang[1], Lei Fu[1], Jie Ren[2], Zhaoxia Yu[3], Ting Chen[2,4,5] and Fengzhu Sun[2,6]**

[1] Department of Automation, Xiamen University, Xiamen, China, [2] Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA, United States, [3] Department of Statistics, University of California, Irvine, Irvine, CA, United States, [4] Bioinformatics Division, Tsinghua National Laboratory of Information Science and Technology, Tsinghua University, Beijing, China, [5] Department of Computer Science and Technology, Tsinghua University, Beijing, China, [6] Center for Computational Systems Biology, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
The University of Texas MD Anderson
Cancer Center, United States

Reviewed by:
Wenxuan Zhong,
University of Georgia, United States
Jonathan Badger,
National Cancer Institute (NCI),
United States

***Correspondence:**
Ying Wang
wangying@xmu.edu.cn
Fengzhu Sun
fsun@usc.edu;
fsun@dornsife.usc.edu

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

Received: 15 November 2017
Accepted: 16 April 2018
Published: 03 May 2018

Citation:
Wang Y, Fu L, Ren J, Yu Z, Chen T
and Sun F (2018) Identifying
Group-Specific Sequences
for Microbial Communities Using
Long k-mer Sequence Signatures.
Front. Microbiol. 9:872.
doi: 10.3389/fmicb.2018.00872

Comparing metagenomic samples is crucial for understanding microbial communities. For different groups of microbial communities, such as human gut metagenomic samples from patients with a certain disease and healthy controls, identifying *group-specific* sequences offers essential information for potential biomarker discovery. A sequence that is present, or rich, in one group, but absent, or scarce, in another group is considered "*group-specific*" in our study. Our main purpose is to discover *group-specific* sequence regions between control and case groups as disease-associated markers. We developed a long *k*-mer ($k \geq 30$ bps)-based computational pipeline to detect *group-specific* sequences at strain resolution free from reference sequences, sequence alignments, and metagenome-wide *de novo* assembly. We called our method MetaGO: *Group-specific* oligonucleotide analysis for metagenomic samples. An open-source pipeline on *Apache Spark* was developed with parallel computing. We applied MetaGO to one simulated and three real metagenomic datasets to evaluate the discriminative capability of identified *group-specific* markers. In the simulated dataset, 99.11% of *group-specific* logical *40*-mers covered 98.89% *disease-specific* regions from the disease-associated strain. In addition, 97.90% of *group-specific* numerical *40*-mers covered 99.61 and 96.39% of differentially abundant genome and regions between two groups, respectively. For a large-scale metagenomic liver cirrhosis (LC)-associated dataset, we identified 37,647 *group-specific 40*-mer features. Any one of the features can predict disease status of the training samples with the average of sensitivity and specificity higher than 0.8. The random forests classification using the top 10 *group-specific* features yielded a higher AUC (from ~0.8 to ~0.9) than that of previous studies. All *group-specific 40*-mers were present in LC patients, but not healthy controls. All the assembled 11 *LC-specific* sequences can be mapped to two strains of *Veillonella parvula*: UTDB1-3 and DSM2008. The experiments on the other two real datasets related to Inflammatory Bowel Disease and Type 2 Diabetes in Women consistently demonstrated that MetaGO achieved better prediction accuracy with fewer features compared to previous studies. The experiments showed that MetaGO is a powerful tool for identifying *group-specific k*-mers, which would be clinically applicable for disease prediction. MetaGO is available at https://github.com/VVsmileyx/MetaGO.

Keywords: long *k*-mer, classification, *group-specific* sequence, metagenomics, microbial community, disease prediction

INTRODUCTION

High-throughput sequencing technologies have ushered in new views of ubiquity and diversity of microbial communities (Yatsunenko et al., 2012). Metagenomic sequencing data permit comprehensive profiling of microbial communities at single-nucleotide resolution. The ability to compare two groups of metagenomic samples is crucial for understanding microbial communities and their effects on hosts. Typically, for two groups of individuals, patients with a certain disease and healthy individuals, *group-specific* markers offer significant support in understanding and predicting disease. Here, "*group-specific* markers" can be genes, species, or sequences present, or rich, in one group, but absent, or scarce, in another group. "*Group-specific*" focuses on the highest discriminative power, rather than the statistically significant difference (White et al., 2009; Segata et al., 2011), to classify, or predict, case and control groups. Accordingly, prediction performance evaluates the discriminative capability of identified *group-specific* features.

Some studies characterized microbiomes by aligning reads to reference genomes or 16S rRNA marker genes (Costello et al., 2009; Quast et al., 2012; Lozupone et al., 2013; Jiang, 2015). It was realized that the alignment-based methods were limited by incomplete or inaccurate reference sequences (Kunin et al., 2008). For example, only about 31.0–48.8% of the shotgun reads from human gut could be aligned to 194 public human gut bacterial genomes, and 7.6–21.2% to the bacterial genomes deposited in GenBank (Qin et al., 2010). Recently, more studies adopted reference-free strategies to analyze the compositional differences of metagenomes between control and case groups at the microbial gene, gene set, or species levels. Generally, contigs were produced through the metagenome-wide *de novo* assembly, and a gene catalog was established through open-reading frame (ORF) prediction. The above processing was first applied to human microbiome of inflammatory bowel disease (IBD) (Qin et al., 2010). Follow-up investigations were conducted based on the constructed gene sets: approximately 60,000 associated gene markers were identified to predict Type 2 Diabetes (T2D), and the concept of a metagenomic linkage group was proposed, which is a group of genes that co-exist among samples and has a consistent abundance level and taxonomic assignments (Qin et al., 2012). The metagenomic gene clusters based on high abundance correlations were further applied to predict T2D in European women using gut metagenomic samples (Karlsson et al., 2013). The gene clusters containing a large number of genes (such as >700) assist *de novo* genome assembly to discover microbial species associated with liver cirrhosis (LC) (Qin et al., 2014) and IBD (Nielsen et al., 2014). Pasolli et al. (2016, 2017) conducted prediction tasks on 2424 metagenomic samples from eight large-scale projects using species-level relative abundances and the presence of strain-specific markers as features. Wen et al. (2017) compared the predicting performances of three types of biomarkers: sequenced reference genomes, genes and gene clusters, for ankylosing spondylitis based on gut metagenomic samples. They found that gene markers performed better than reference genome markers and clustered gene markers, and the clustered gene markers might be limited by the unknown taxonomic organisms in clusters. Almost all the above studies followed the analysis pipeline of *de novo* contig assembly, gene prediction, and gene clustering. Previous studies concluded that metagenome-assembly performs well for microbial communities that have high coverage of phylogenetically distinct, and low taxonomic diversity (Papudeshi et al., 2017), but the presence of closely related strains in one community would substantially have negative effect on the assembly performance (Sangwan et al., 2016; Sczyrba et al., 2017). Moreover, high co-abundance among species would result in multiple species in one cluster (Nielsen et al., 2014). Therefore, components with closely related genome sequences or abundance would diminish the performance of assembly and clustering in microbial community studies.

Besides genes or species, assembled contigs have also been used as features to predict disease. Several contig binning tools, such as CONCOCT (Alneberg et al., 2014), MaxBin2.0 (Wu et al., 2016), COCACOLA (Lu et al., 2017), and MetaGen (Xing et al., 2017), were developed for binning contigs assuming that contigs with similar coverage/relative abundances over different samples come from the same genomes. In particular, although the main purpose of MetaGen (Xing et al., 2017) is to identify microbial species in the community through binning, the study not only designed comprehensive experiments to analyze the effect of sequencing depth, sample size, number of species and sequence similarity, but also used the relative abundance of each bin to predict IBD/T2D/obesity disease on metagenomic datasets to evaluate the binned microbial composition. Similarly, Ren et al. (2017) developed a novel pipeline to predict the disease status of LC using the abundance of viral contig bins. Both studies made novel attempts to identify markers through assembling *de novo* reads into contigs and then binning contigs, which achieved excellent predicting results. The basic idea is to discover species markers that are differentially abundant between case and control groups. However, current assembly tools are hard to handle large-scale datasets: reads assembly involves the construction of *De Bruijn* graph, error correction, and path resolution; contig binning requires mapping reads to the assembled contigs; both would require extremely large memory and are very time-consuming. Also, if the main purpose is to discover *group-specific* markers, it is not necessary to assemble contigs for the genomes that are not associated with the disease.

The *k*-mer frequencies (i.e., the number of occurrences of *k*-mers within the whole sequencing data) are another representative alignment-free feature to characterize a microbial community. The frequency distributions of *2–10*-mers were used to compare metagenomic and meta-transcriptomic communities (Jiang et al., 2012; Wang et al., 2014; Liao et al., 2016) or to improve contig binning within a community (Wang et al., 2017). Also, Cui and Zhang (2013) classified clinical metagenomic samples using the frequencies of *2–10*-mers.

However, *2–10*-mers are too short to capture specific details inside the microbial community, such as sequences present, or rich, in one group, but absent, or scarce in another group. Intuitively, longer *k*-mers contain richer biological information in the nucleotide sequences. The long *k*-mers had been mainly utilized as seed index in sequence assembly and alignment (Li et al., 2010; Grabherr et al., 2011). Recently, long *k*-mers ($\geq$20 bp) began to be utilized to more applications: our previous study explored the effect of *k*-mer length on an unsupervised comparison between metagenomic samples and verified the promising performance of long *k*-mers to depict the specific characteristics of microbial communities (Wang et al., 2015). Han et al. (2017) detected differentially abundant *21*-mers in metagenomic samples from T2D and healthy individuals, assembled the reads containing those *21*-mers into contigs, and then predicted genes based on the contigs. Finally, they used the gene abundances to predict T2D status. Our study differs from Han et al. (2017) in the sense that we do not predict genes based on the contigs assembled from reads containing statistically differentially abundant *k*-mers. Instead, we identified *group-specific k*-mers using discriminative power to separate two groups and predicted disease status with *k*-mers as features. Moreover, *group-specific k*-mers were assembled to contigs directly. Rahman et al. (unpublished) found significant differentially abundant *31*-mers between two groups of 1000 genomes data and discovered SNPs between different populations, which is highly different from the objectives of this study. The frequency vector of long *k*-mers ($\sim$30 bp) was also applied to calculate the dissimilarity between metagenomic samples using 16 standard ecological distances (Benoit et al., 2016). The long *k*-mers began to present attractive potentials to characterize high-throughput sequencing data.

Since sufficiently long *k*-mers are usually specific to a genome (Fofanov et al., 2004), therefore, we proposed a computational framework to identify *group-specific* sequences between two groups of metagenomic samples with long ($\geq$30 bp) *k*-mers in this study. We call our method MetaGO: *Group-specific* oligonucleotide analysis for metagenomic samples. The main purpose of MetaGO is to discover *group-specific* sequence regions between control and case groups as disease-associated markers. Instead of using statistically significant difference as index, we considered the discriminant power to separate two groups of single *k*-mer. A *k*-mer is considered *group-specific* if (1) the average of sensitivity and specificity (ASS) is higher than a preset threshold when using the presence/absence of the *k*-mer on the sequencing data to predict disease status, or (2) the *k*-mer's frequencies are significantly different between two groups of samples (Wilcoxon rank-sum test, *p*-value $\leq$ 0.01) and the ASS is higher than a preset threshold using logistic regression. The *group-specific k*-mers are identified based on the training set. In our study, *k*-mer length is set between 30 and 40 given the tradeoff among sensitivity, specificity, and computational cost. To reduce the computational burden from long *k*-mers, we developed an open-source, parallel-computing pipeline on *Apache Spark*. Once the *group-specific k*-mers are identified, we assembled them into *group-specific* sequences. The assembly on the markedly reduced number of long *k*-mers will be more computationally efficient and accurate.

MetaGO was tested on one simulated and three real metagenomic datasets. In the simulated dataset, for the two strains sharing 87% common sequences where one is disease specific and the other one is present in both groups, we identified *group-specific* logical *40*-mers that covered 98.89% (recall) of the *disease-specific* sequence regions from the disease-associated strain with 98.91% precision. In addition, 98.83% of the *group-specific* numerical *40*-mers covered 99.01 and 97.30% of the differential-abundant genome and regions, respectively. For the metagenomic LC-associated dataset (Qin et al., 2014), it is composed of human fecal samples from 98 LC patients and 83 healthy controls, as well as an additional independent dataset containing 25 patients and 31 controls. The *k*-mer length was set as 40 because of the large sample size (number of samples). In our experiment, two-thirds of the 98 patients and 83 control samples were randomly selected as the training set, leaving one-third as the validation set and the extra 25 patients and 31 controls as the independent testing set. In total, 37,647 *group-specific 40*-mers were identified on the training set, and 35,652 and 12,944 of the *group-specific 40*-mers yielded ASS $\geq$ 0.8 on the validation and testing sets, respectively. The *single-logical-feature* predictor with the highest ASS score 0.87 on the training set predicted the disease status in the validation and testing sets with ASS score as 0.88 and 0.83, respectively. Using the top 10 *group-specific 40*-mers, the random forests classifier achieved the area under the receiver operating characteristic (AUC) as 0.963, 0.969, and 0.942 on training, validation, and testing sets, respectively. It is interesting to note that all 37,647 *40*-mers were present in LC patients but absent from healthy controls. The *LC-specific 40*-mers were assembled into 11 sequences with a length between 210 and 350 bp, and they demonstrated the distinguishing coverages between two groups. All the identified *LC-specific* sequences could be matched to two strains of *Veillonella parvula*, UTDB1-3 and DSM2008 with 97–100% identity. And 83.2 and 79.6% of the 37,647 *group-specific 40*-mers could be matched to strain UTDB1-3 and DSM2008, respectively.

We also identified *group-specific* *k*-mers based on two more metagenomic disease-associated datasets: IBD associated (Qin et al., 2010) and WT2D (T2D in women) associated (Karlsson et al., 2013). Based on the identified *group-specific* *k*-mers, our pipeline achieved substantially better prediction performance using relatively fewer features compared with previous studies having identical or relaxed experimental settings. All experiments demonstrated long *k*-mers to be more efficient in capturing the specific information of sequencing data and discriminating gut microbiome communities between control and case groups. It should be noted that *group-specific* sequences are identified free from reference sequences, metagenome-wide assembly, and sequence alignments. MetaGO greatly facilitates the identification of clinically meaningful biomarkers.

MATERIALS AND METHODS

Description of Terms

A group-specific feature is a *k*-mer present, or rich, in the metagenomic sequencing data of one group, but absent, or sparse, in the sequencing data of another group. A *k*-mer is a word composed of *k* oligonucleotides, and the total number of all possible *k*-mers is 4^k.

We defined *k*-mer features in the following two ways:

Numerical features are the normalized frequencies of *k*-mers. The numerical feature of a *k*-mer *i* in sample *j* is denoted as $f_i(j)$ and is defined in Equation (1), where $f_i^{\circ}(j)$ is the number of occurrences of *k*-mer *i* in sample *j*, and *n* is the total number of *k*-mers, that is 4^k. So the normalization is the number of occurrences of the *k*-mer over the total number of occurrences for all *k*-mers in one sample. Each *k*-mer has the same length *k*, so length is not considered during the normalization.

$$f_i(j) = \frac{f_i^{\circ}(j)}{\sum_{i=1}^{n} f_i^{\circ}(j)}, \quad i = 1, 2, \ldots, n. \tag{1}$$

Logical features are the logicalization of numerical features. They use 1 and 0 to represent *k*-mers as present or absent in one sample, as shown in Equation (2),

$$f_i^{(l)}(j) = \begin{cases} 1 & \text{if } f_i(j) > 0 \\ 0 & \text{if } f_i(j) = 0 \end{cases}, \tag{2}$$

where $f_i^{(l)}(j)$ is the logical value of *k*-mer *i* in sample *j*, and the superscript "*l*" indicates logical feature.

A single-logical-feature predictor, as represented in Equations (3) and (4), is used to predict disease status based on whether a *k*-mer *i* is present in the sequencing data of sample *j* or not.

$$f_i^{(l)}(j) = \begin{cases} 1 & \text{then sample } j \in \text{Group} + \\ 0 & \text{then sample } j \in \text{Group} - \end{cases} \tag{3}$$

or

$$f_i^{(l)}(j) = \begin{cases} 1 & \text{then sample } j \in \text{Group} - \\ 0 & \text{then sample } j \in \text{Group} + \end{cases}. \tag{4}$$

A single-numerical-feature logistic regression predicts the case and control status based on one single numerical feature, and it is used as the independent variable in a logistic regression. An example of each term above is given in **Supplementary File S1**.

The Computational Framework to Identify *Group-Specific* Sequences

As shown in **Figure 1**, the computational framework of MetaGO consists of three modules. (1) *Creating a feature vector for each sample*. The feature vector is composed of the number of occurrences for each *k*-mer through all reads in one sample. (2) *Feature preprocessing*. After removing *k*-mers occurring only once and normalizing *k*-mer frequencies, the feature matrix is integrated on the feature vectors across all training samples. The *k*-mers that are absent in most training samples are filtered out. (3) *Identifying group-specific features*. The logical and numerical features with high discriminant power are selected.

MetaGO was developed on *Apache Spark* to reduce computational costs through parallel running on HDFS of Hadoop or a stand-alone multi-core server. The open-source pipeline is available at https://github.com/VVsmileyx/MetaGO.

Module 1: Creating Feature Vectors

A feature vector consists of elements that account for the number of occurrences (i.e., frequency) for each *k*-mer through all the reads in one metagenomic sample. Existing tools, such as DSK (Rizk et al., 2013) or JELLYFISH (Marçais and Kingsford, 2011), are available for counting *k*-mer frequency. In our study, we used DSK to count *k*-mers. The reverse complements of reads were taken into consideration. A *k*-mer and its reverse complement were considered as the same object, so the theoretical dimension of a feature vector for one sample is shrunk to $\frac{4^k+2^k}{2}$ for even *k* and $\frac{4^k}{2}$ for odd *k*. Furthermore, only the *k*-mers that occur in a sample are stored in the feature vector to reduce storage space.

Module 2: Feature Preprocessing

Discard *k*-mer Features Occurring Only Once

With the increase of *k*-mer length, *k*-mer frequency decreases exponentially, and the *k*-mer vector is highly sparse. A *k*-mer occurring only once might be caused by low abundance or sequencing errors. To achieve reproducible and stable prediction models, *k*-mers occurring once were removed from the frequency vector, and this step was implemented by DSK during *k*-mer counting in our study.

Normalize *k*-mer Frequencies

Owing to different sequencing depths in samples, the frequency of a *k*-mer is normalized using Equation (1) by the total number of occurrences of all *k*-mers.

Build Feature Matrix Across Training Samples

Feature vectors across all training samples are integrated as a matrix. This step is extremely time- and memory-consuming as a result of the large sample size and the long *k*-mer length. Just storing non-zero *k*-mers in each feature vector, the integration process requires huge amounts of sorting and matching of

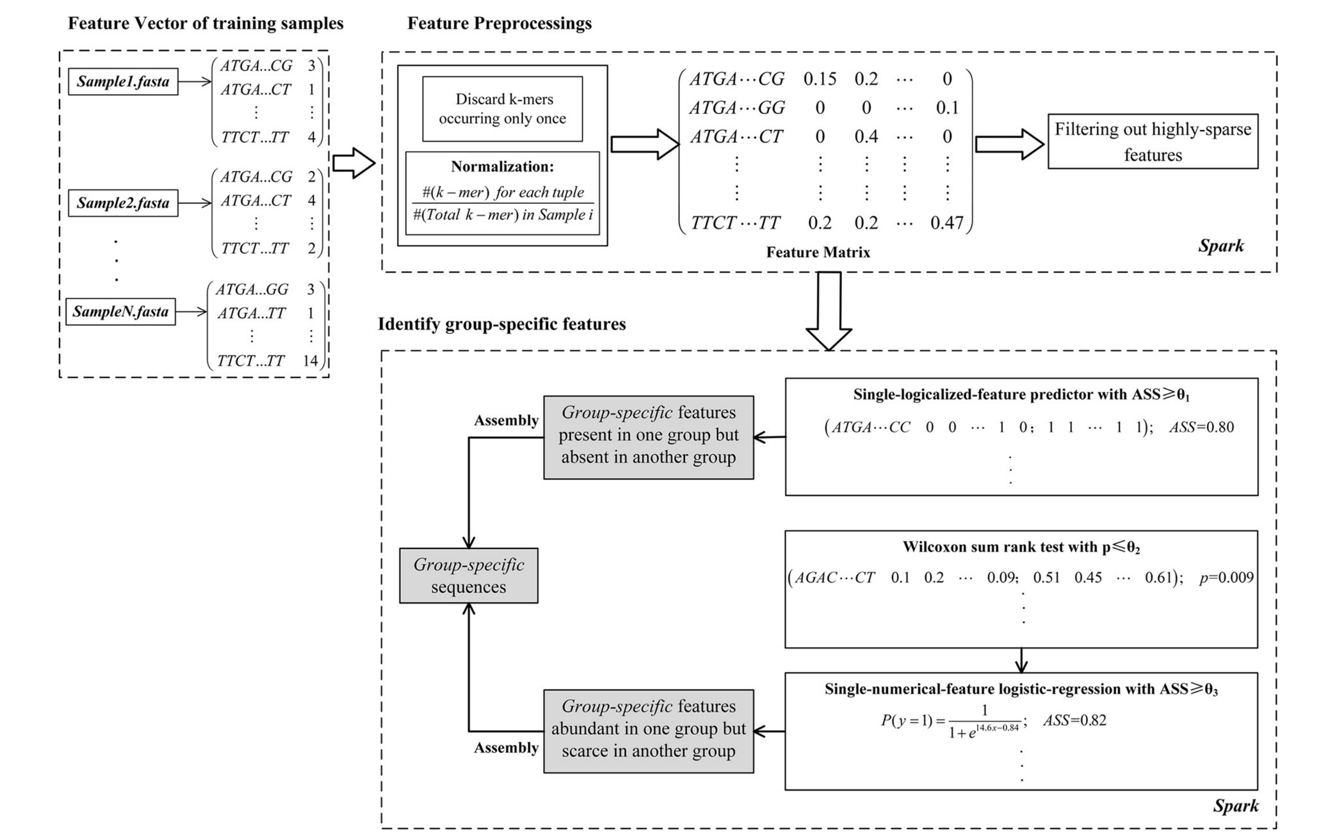

Feature Vector of training samples
Feature Preprocessings
Sample1.fasta
ATGA...CG 3
ATGA...CT 1
TTCT...TT 4
Sample2.fasta
ATGA...CG 2
ATGA...CT 4
TTCT...TT 2
SampleN.fasta
ATGA...GG 3
ATGA...TT 1
TTCT...TT 14
Discard k-mers occurring only once
Normalization:
#(k – mer) for each tuple / #(Total k – mer) in Sample i
ATGA···CG 0.15 0.2 ··· 0
ATGA···GG 0 0 ··· 0.1
ATGA···CT 0 0.4 ··· 0
TTCT···TT 0.2 0.2 ··· 0.47
Feature Matrix
Filtering out highly-sparse features
Spark
Identify group-specific features
Single-logicalized-feature predictor with ASS≥θ1
(ATGA···CC 0 0 ··· 1 0; 1 1 ··· 1 1); ASS=0.80
Group-specific features present in one group but absent in another group
Assembly
Group-specific sequences
Wilcoxon sum rank test with p≤θ2
(AGAC···CT 0.1 0.2 ··· 0.09; 0.51 0.45 ··· 0.61); p=0.009
Single-numerical-feature logistic-regression with ASS≥θ3
P(y = 1) = 1 / (1 + e^(14.6x−0.84)); ASS=0.82
Group-specific features abundant in one group but scarce in another group
Assembly
Spark

FIGURE 1 | The MetaGO framework to identify *group-specific* sequences with long *k*-mer features. The framework is composed of three modules. (1) The feature vector of each metagenomic sample is composed of the frequencies of all *k*-mers. (2) The *k*-mers are preprocessed by discarding features occurring only once, normalization, integrating the matrix and removing the *k*-mers absent from most training samples. (3) The features are represented as logical and numerical forms, and the features with high discriminant power are identified to be *group-specific*.

k-mers. When $k = 40$, approximately 10^9 *40*-mer features occur more than once. The feature matrix F is denoted as Equation (5), where $k\text{-mer}_1, k\text{-mer}_2, \ldots, k\text{-mer}_m$ are the m *k*-mer features, and $S_1, S_2, \ldots, S_N$ are the N training samples from case and control groups.

$$F = \begin{array}{c} \\ k{-}\mathrm{mer}_1 \\ k{-}\mathrm{mer}_2 \\ \vdots \\ k{-}\mathrm{mer}_m \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & \cdots & S_N \end{array} \\ \begin{pmatrix} f_1(1) & f_1(2) & \cdots & f_1(N) \\ f_2(1) & f_2(2) & \cdots & f_2(N) \\ \vdots & \vdots & \vdots & \vdots \\ f_m(1) & f_m(2) & \cdots & f_m(N) \end{pmatrix} \end{array} \tag{5}$$

Remove Highly-Sparse Features

The "highly-sparse" feature means that a *k*-mer is absent in most training samples, i.e., the frequencies of *k*-mers are 0 in most training cases and controls. Such features have limited contributions to classification. In our study, if a *k*-mer is absent in more than 80% of control samples and 80% of case samples, the feature is removed. The stringent threshold of 80% offers high confidence in filtering out less useful features.

Module 3: Identifying *Group-Specific* Features

After preprocessing, about 10^6 features still remain for *40*-mers. Simple feature-ranking filtering is more suitable than Wrapper feature selection. Wrapper methods consider the selection of a set of features as a search problem in which different combinations are prepared, evaluated, and compared to other combinations. The dimension of combination space is extremely high for a large number of features in our study. The filtering of *k*-mers is only based on the training data without touching the validation and testing data.

Identify *Group-Specific* Logical Features Based on a *Single-Logical-Feature* Predictor

As shown in **Figure 2**, numerical features were transformed to logical features using Equation (2), and the *single-logical-feature* predictors were created according to Equations (3) or (4). The performance of a predictor was evaluated by ASS, an average of sensitivity and specificity. If a *single-logical-feature* predictor achieves ASS $\geq \theta_1$, the corresponding *k*-mer is identified to be *group specific*. The *group-specific* logical features are present in one group but absent in another group.

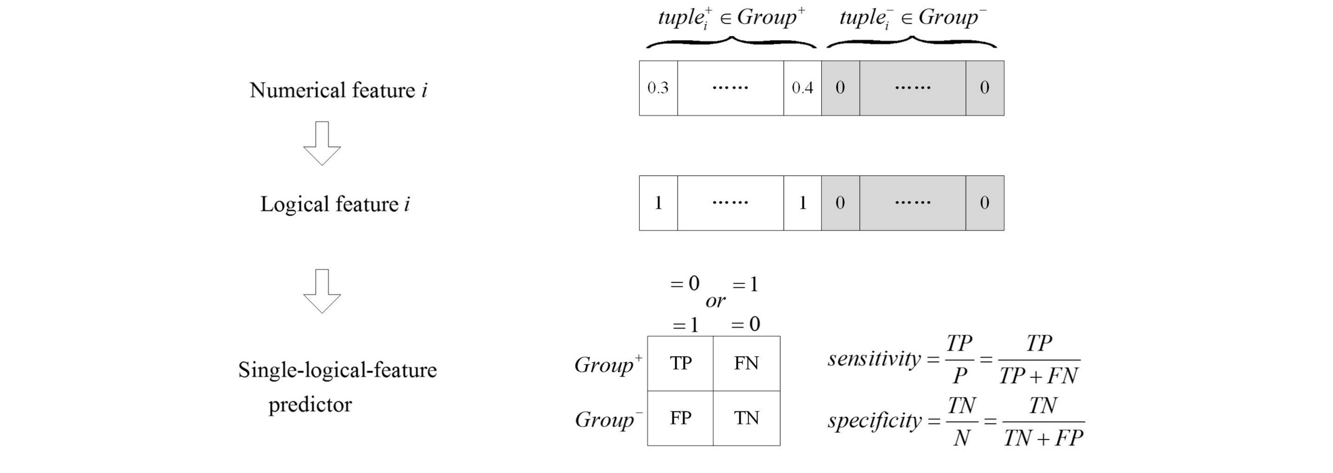


FIGURE 2 | The *single-logical-feature* predictor. The numerical feature is transformed into the logical feature. Based on the logical value of the feature, the *single-logical-feature* predictor is designed, and the corresponding ASS is calculated.

In our study, θ_1 was set as 0.80, which means that each *group-specific k*-mer alone can separate two groups of training samples with ASS $\geq$ 0.8 solely. Some researchers would prefer a statistical test, such as Chi-squared test, to rank the features. To accommodate this preference, we calculated *p*-values of Chi-squared test for the same feature set. Among the two feature lists with the 400 largest ASS values and the 400 smallest *p*-values, 392 features were present in both lists in the same order. Therefore, both ASS and Chi-squared test provide consistent ranks of the features. In our pipeline, users have the option to choose either ASS or Chi-squared test as evaluation metrics.

Identify *Group-Specific* Numerical Features Based on a *Single-Numerical-Feature* Logistic-Regression Predictor

First, Wilcoxon rank-sum test is applied to the numerical features to select *k*-mers with differential abundance (*p*-value $\leq \theta_2$) between two groups. However, our main goal is to identify features with the most discriminant power. Therefore, we fit logistic regression for each numerical *k*-mer feature that passed the Wilcoxon rank-sum test over all the training samples, and we term this as *single-numerical-feature* logistic-regression predictor. We used ASS $\geq \theta_3$ as a metric to identify *group-specific* numerical *k*-mers. In our study, we used $\theta_2 = 0.01$ and $\theta_3 = 0.8$

Random Forests Prediction of Disease Status With the Combination of Multiple Features

The *single-logical-feature* predictor and *single-numerical* logistic-regression predictor are the classifiers based on a single *k*-mer feature. Because of the complicated association between human microbiome and disease, classifiers using multiple features are expected to be more efficient than those with single features. Therefore, we used random forests to design a classifier with multiple *group-specific* features. To remove redundant features, we calculated the Pearson correlation coefficients (PCC) between the feature vectors of every pair of *k*-mers. If a pair of *k*-mers has a PCC value higher than a preset threshold, such as 0.75, one *k*-mer feature was randomly discarded. The remaining features were ranked according to the variable importance measures of Breiman's random forests method (Breiman, 2001), and the top features were adopted to design a random forests classifier.

Assembly of *Group-Specific* Sequences

Using CAP3 (Huang and Madan, 1999), the identified *group-specific k*-mers based on logical and numerical features were, respectively, assembled to longer sequences. For quality control, the assembled sequences longer than a certain threshold (200 bp in our study) are considered as *group-specific* sequences.

Parallel Computing Workflow on *Apache Spark*

The running time and memory required to integrate feature matrix and filter out less useful features expand dramatically with the increase of *k*-mer length and sample size. Fortunately, these processing steps are suitable for parallel computing. Therefore, we developed MetaGO workflow on *Apache Spark* (Zaharia et al., 2010) to implement parallel computing. *Spark* can run in local mode or cluster mode. Thus, MetaGO can run on a local stand-alone multi-core server or a distributed cluster on HDFS. The detailed description of the workflow is given in **Supplementary File S1**. The workflow is available on https://github.com/VVsmileyx/MetaGO.

Experimental Design

The Setting of *k*-mer Length

A previous study showed that sufficiently long *k*-mers are usually specific to a genome (Fofanov et al., 2004). According to an observation based on 100 pairs of bacterial genomes, the average ratio of common *k*-mers between the genomes is <1.02% when $k \geq 30$ (Le et al., 2015). Therefore, *k*-mers longer than 30 bp would possess sufficiently high sensitivity to capture the discriminate characteristics to separate two groups; thus, theoretically, longer *k*-mers are better suited to this task.

At the same time, however, *k*-mer length is limited by four factors: sample size (the number of samples), sequencing depth, computational cost, and read length. First, the dimension of feature space grows exponentially with *k*. Owing to the curse of dimensionality, a limited number of samples would lead to a high false-positive rate. Therefore, a large sample size is required to obtain high specificity. Second, when sequencing depth is not deep enough to cover all the metagenomic regions, the frequencies of long *k*-mers would not be accurate. Third, with the increase of *k*-mer length, the huge number of *k*-mers leads to the explosion of memory and storage. Fourth, when the *k*-mer length is close to read length, the frequencies of *k*-mers are contaminated by the truncated sites under limited sequencing depth. Therefore, we set the *k*-mer length to be 30–40 as the reasonable tradeoff among sensitivity, specificity, and computational cost.

Simulated Metagenomic Dataset

Based on the relative abundances of frequent microbial genomes within human gut analyzed by Qin et al. (2010) (Figure 3 of their paper), we selected the top 10 most frequent genomes as the basis components of the simulation. The relative abundances in the control group were approximated from the medians of Figure 3 of that study (Qin et al., 2010), which were converted into the cell proportions of the 10 genomes in all the cells within the community. In addition, we added another strain *Bacteroides thetaiotaomicron* VPI-5482 to the patient group, and this strain shares about 87% common sequences with the existing *B. thetaiotaomicron* 7330. Meanwhile, we assigned Genome *Bacteroides caccae* ATCC 43185 threefold abundance in the control group than in the patient group. The remaining nine genomes have identical abundance distributions between the healthy individual and the patient groups. The detail setting is shown in **Table 1**. We used MetaSim (Richter et al., 2008) to generate 15 metagenomic samples for case and control groups, respectively. For each group, the absolute values of Gaussian noises of mean zero and standard derivation equal to each central relative abundance were added to the center relative abundance vector. Each sample contains ∼10,000,000 reads. In the evaluations, the proportion of identified *group-specific k*-mers that can be aligned to disease-specific sequence regions is called "precision," and the proportion of disease-specific sequence regions that can be covered by *group-specific 40*-mers is called "recall."

Metagenomic Liver Cirrhosis-Associated Dataset

In recent studies, alterations in human gut microbiota have been linked to LC (Qin et al., 2014; Wiest et al., 2014). We analyzed the human fecal metagenomic samples (Qin et al., 2014) from 98 LC patients and 83 healthy controls, as well as an extra dataset composed of 25 independent patients and 31 controls. The data were sequenced with Illumina HiSeq 2000. In the experiment, two-thirds of the 98 patients and 83 control samples were randomly selected as the training set to identify *group-specific k*-mers, and the remaining one-third as the validation set. Finally, the extra 25 patients and 31 controls were applied to test the *group-specific k*-mers independently.

Metagenomic IBD-Associated and WT2D-Associated Datasets

The IBD dataset is composed of the human fecal metagenomic samples from 25 IBD patients and 97 controls (Qin et al., 2010). These samples were sequenced on Illumina GAIIx from the MetaHIT project (Human Microbiome Project Consortium, 2012). The WT2D dataset is composed of samples from 53 T2D patients and 43 healthy controls from European women (Karlsson et al., 2013). These samples were sequenced on Illumina HiSeq 2000. Both datasets had been predicted using various types of features (Cui and Zhang, 2013; Karlsson et al., 2013; Pasolli et al., 2016). In our study, we adopted the experimental setting of a previous study (Pasolli et al., 2016), in which 20 independent runs of 10-fold cross-validation were used to evaluate the classification.

RESULTS

The Simulated Metagenomic Dataset

For logical features, there were 1,646,128 *group-specific 40*-mers using ASS $\geq$ 0.8 as a threshold. And 99.999% of the *40*-mers were patient specific, which means almost all the logical group-specific *40*-mers exist only in the patient group and are absent in the healthy control group. Among the logical *patient-specific 40*-mers, 99.11% of them (precision) were exactly aligned to strain *B. thetaiotaomicron* VPI-5482 (the strain present in the patient group only) and covered 98.89% (recall) of the regions that are not in the genome of the other strain *B. thetaiotaomicron* 7330. None of the *group-specific* 40-mers were aligned to *B. thetaiotaomicron* 7330, which has the same abundance on both groups. The logical *group-specific 40*-mers mainly indicate genomes present in one group but not in another group.

The remaining features were represented as numerical *40*-mers, and there were 7,891,412 *group-specific 40*-mers using $p < 0.05$ and ASS $\geq$ 0.8 as the thresholds. And 4,452,553 (56.42%) of them were exactly matched to *B. caccae* ATCC 43185 and covered 99.61% (recall) of the whole genome, which is differentially abundant between the healthy control and the case groups. Among the remaining *40*-mers, 3,257,251 (41.3%) of them were aligned to the common regions between *B. thetaiotaomicron* VPI-5482 and *B. thetaiotaomicron* 7330, and covered 96.39% (recall) of the common sequences. Because for the patient group, the abundance of common sequences includes VPI-5482 and *B. thetaiotaomicron* 7330, but the control group only includes *B. thetaiotaomicron* 7330, the common sequences are differentially abundant. In total, 97.72% (precision) of the identified *group-specific* numerical *40*-mers were aligned to the differentially abundant regions between the two groups.

The identified *patient-specific* and *control-specific 40*-mers from logical and numerical features were assembled into contigs, respectively. For the assembled *patient-specific* contigs, there were 20 of them with length ≥10,000 bp and all these contigs were matched to the *patient-specific* strain *B. thetaiotaomicron* VPI-5482 with 99.79–100% identity and 100% coverage. The coverage rate here means the proportion of contig sequence mapped

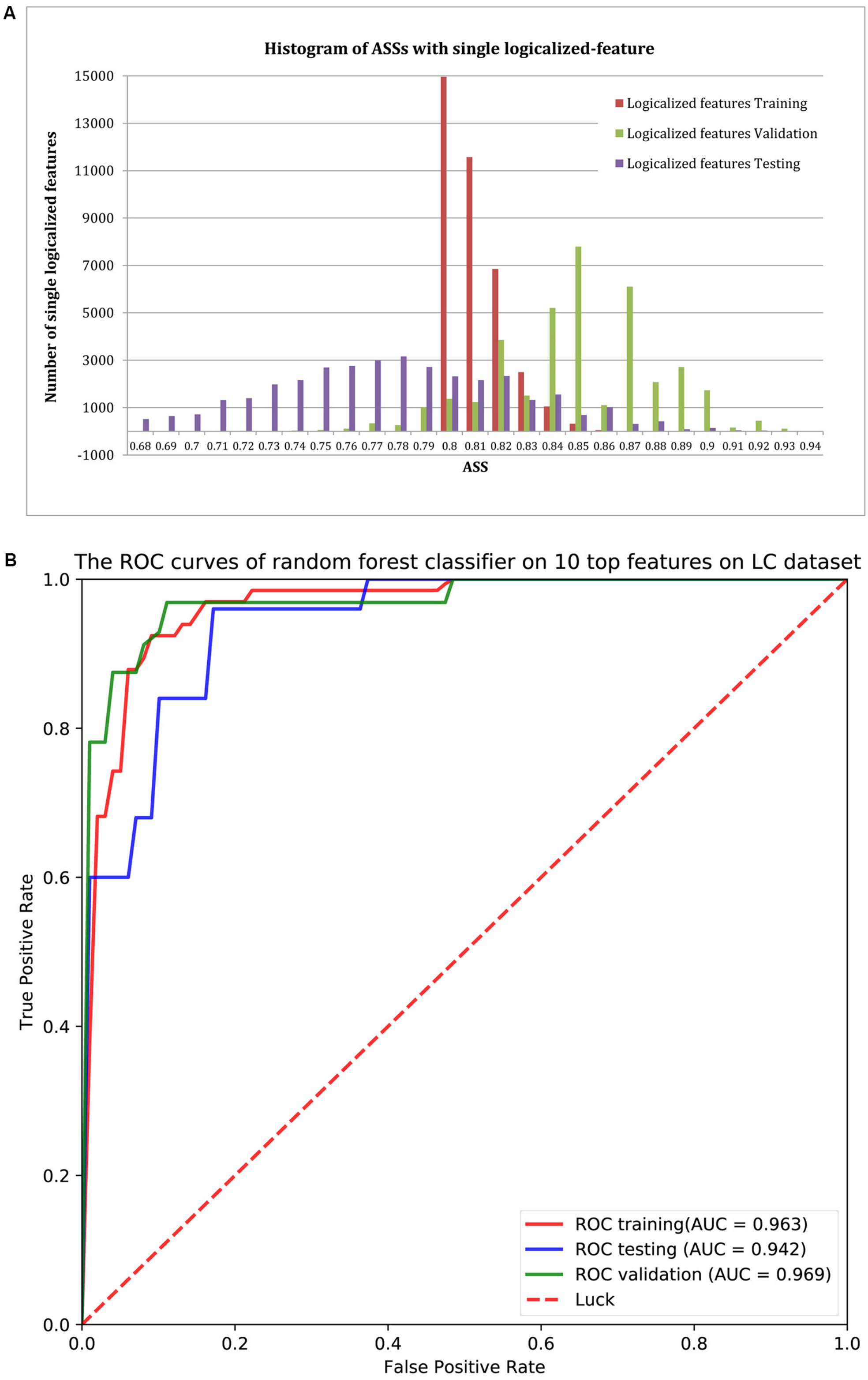


FIGURE 3 | (A) The distribution of ASS values of the 37,302 *single-logical-feature* predictors and 345 *single-numerical* logistic-regression predictors on the identified *group-specific* features for training, validation, and testing sets. These predictors achieved better performance in the validation set compared to the training set. A total of 35,652 *group-specific* features achieved ASS ≥ 0.8 for the validation set, and 12,944 of them achieved ASS ≥ 0.8 for the testing set. **(B)** ROC curves of the random forests classifier with the top 10 features on validation and testing sets. Using the top 10 *group-specific 40*-mers, the random forests classifier achieved AUC of 0.963, 0.969, and 0.942 on training, validation, and testing sets, respectively.

TABLE 1 | The relative abundance profile of different genomes in control and patient groups for the simulated dataset.

Genomes	NCBI Accession ID	Relative_Abundance_H*	Relative_Abundance_P*
Bacteroides thetaiotaomicron 7330	NZ_CP012937.1	18%	
Bacteroides thetaiotaomicron VPI-5482	NC_004663.1	0	6%
Bacteroides uniformis CL03T12C37	NZ_JH724268.1	7%	
Alistipes putredinis isolate CAG	MNQH01000001.1	16%	
Parabacteroides merdae 2789STDY5834848	CZAG01000002.1	10%	
Dorea longicatena 2789STDY5834914	NZ_CZAY01000001.1	10%	
Ruminococcus bromii L2-63	FP929051.1	10%	
Bacteroides caccae ATCC 43185	NZ_CP022412.2	9%	3%
Clostridium sp. SS2/1	NZ_DS547029.1	8%	
Eubacterium hallii isolate EH1	NZ_LT907978.1	6%	
Ruminococcus torques L2-14	FP929055.1	6%	

*The relative abundances were the proportions of the number of copies of 11 genomes within the community. Bacteroides thetaiotaomicron VPI-5482 is present only in the patient group, and it is another strain of B. thetaiotaomicron. Bacteroides caccae ATCC 43185 has threefold abundance in the control group of that in the patient group. *H, healthy control; P, patient.*

to the strain. In contrast, these contigs cannot be matched to *B. thetaiotaomicron* 7330, and the maximum common sequences between contigs and *B. thetaiotaomicron* 7330 genome were no longer than 47 bp. For assembled *control-specific* contigs, there were 24 of them with length ≥5000 bp and all of them were mapped to the differentially abundant genome *B. caccae* with 100% identity and 100% coverage using BLAST (Altschul et al., 1997).

To evaluate the effect of *k*-mer length, we ran MetaGO on *10*-mer, *20*-mer, *30*-mer, *50*-mer, and *60*-mer, and the corresponding precision and recall are shown in **Table 2**. For the simulated dataset, When $k = 10$, no *group-specific* logical *k*-mers were identified. The recall rates for the identified numerical *k*-mers were only 25.34% for *B. caccae* ATCC 43185 and 22.45% for the common regions between *B. thetaiotaomicron* VPI-5482 and *B. thetaiotaomicron* 7330. When $k \geq 20$, the effects of the *k*-mer length on the performance of our methods were small. The precision increased slightly with the *k*-mer length from 99.03 to 99.35% for logical *k*-mers and from 96.81 to 98.58% for numerical *k*-mers, consistent with the intuition that long *k*-mers can capture more specific information of each group. On the other hand, though almost all the recall rates were all above 90%, the recall first increased with *k*-mer length until $k = 40$ and then decreased, which might be caused by insufficient coverage for long *k*-mers.

The experimental results demonstrate that the identification of *group-specific 40*-mers can not only capture genomes with different abundance but also identify *group-specific* markers under the strain-level resolution. Even though the two strains *B. thetaiotaomicron* VPI-5482 and *B. thetaiotaomicron* 7330 share 87% common sequences, our method still captured the *group-specific* sequences.

The LC-Associated Metagenomic Dataset

MetaGO was applied to the large-scale metagenomic LC-associated dataset (Qin et al., 2014). With sufficient training samples and long read length, the *k*-mer length was set as $k = 40$. A total of $\sim 10^9$ non-zero *40*-mers were found in the feature matrix of training samples. After removing the highly sparse *40*-mer features, $\sim 10^6$ features were left.

Identify *Group-Specific* Features

Using ASS > 0.8 as the threshold, 37,302 logical features were identified as *group-specific 40*-mers. That is, any one of these *40*-mers could achieve ASS > 0.8 using its corresponding *single-logical-feature* predictor on training samples. We then used each of these 37,302 *single-logical-feature* predictors to predict LC in the validation and testing sets. As shown in the histogram of **Figure 3A**, ASS values of validation and testing were centered at 0.85 and 0.78, respectively. Among the 37,302 *single-logical-feature* predictors, 35,404 (95%) *group-specific 40*-mers achieved ASS ≥ 0.8 on the validation set, and 12,750 (36%) achieved ASS ≥ 0.8 on the testing set. Furthermore, 345 numerical features were identified as *group-specific 40*-mers with ASS ≥ 0.8, where 248 and 194, respectively, achieved ASS ≥ 0.8 on validation and testing sets using corresponding *single-numerical-feature* logistic regression predictors. All 37,302 logical and 345 numerical *40*-mers were *LC-specific* in that they were all present only in the fecal samples of LC patients, but not in the samples from healthy controls. The identified *group-specific 40*-mers for the LC dataset are available in **Supplementary File S2**.

We also implemented a controlled trial by shuffling the labels of the training samples randomly. Using the same pipeline and settings, only 247 *40*-mers achieved ASS ≥ 0.7, and the highest value was 0.73. This control trial indicates that most of the identified *group-specific 40*-mers for LC were more likely to be true rather than due to false positives.

Classification With the *Group-Specific* 40-mer(s)

We used classification performance to evaluate the discriminative capability of the identified *group-specific 40*-mers. First, we classified the healthy and LC groups with single features. The *single-logical-feature* predictor that obtained the highest ASS = 0.87 on the training set achieved ASS = 0.885 (sensitivity = 0.81 and specificity = 0.96) on the validation set and 0.87 (sensitivity = 0.84 and specificity = 0.90) on the independent testing set. Second, we built a classifier using a set of features.

TABLE 2 | The precision and recall of MetaGO for the simulated dataset using different *k*-mer lengths.

***k*-mer length**		**10 (%)**	**20 (%)**	**30 (%)**	**40 (%)**	**50 (%)**	**60 (%)**
Logicalized *k*-mers	Precision	–*	99.03	99.05	99.11	99.45	99.35
	Recall	–*	89.79	92.16	98.89	97.01	95.23
Numercial *k*-mers	Precision	99.63	96.81	96.07	97.72	98.22	98.58
	Averaged recall	23.89	95.70	97.93	98.00	96.82	94.76

*The "averaged recall" in numerical k-mers is the average of the recall of B. caccae ATCC 43185 genome and the recall of the common regions between strain B. thetaiotaomicron 7330 and B. thetaiotaomicron VPI-5482. *When k = 10, there is no logicalized k-mer identified, so it is marked with "–".*

TABLE 3 | Comparison of the prediction performance of different methods based on the LC dataset.

Feature		***40*-mer**	***40*-mer**	**Gene markers††**	**Species abundance†**	**Presence of strain-specific markers†**
Experiment		Training (66P+56H) Validation (32P+27H) Testing (25P+31H)			20 runs of 10-fold cross-validation (114P+118H)	
Number of feature		**1**	**10**	15	542	120553
Classifier		**Single logical feature predictor**	**Random forests**	Support vector machine	Random forests	Support vector machine
AUC	Training	**ASS* = 0.87**	**0.963**	0.918	0.946 ± 0.035	0.963 ± 0.027
	validation	**ASS = 0.885**	**0.969**	0.838		
	testing	**ASS = 0.87**	**0.942**	0.836		

*Using much fewer features, MetaGO achieved better results compared to other methods. The results of MetaGO were in bold. †(Pasolli et al., 2016); ††(Qin et al., 2014); *average of sensitivity and specificity.*

Using the top 10 *group-specific 40*-mers, a random forests classifier achieved AUCs of 0.963 on training, 0.969 on validation, and 0.942 on testing sets, respectively. The corresponding ROC curves are shown in **Figure 3B**. As shown in **Table 3**, Qin et al. (2014) obtained AUC = 0.918, 0.838, and 0.836 on training, validation, and testing sets with SVM using 15 marker genes as features. Pasolli et al. (2016) obtained AUC = 0.946 ± 0.036 with random forests using 542 species-abundance features and 0.963 ± 0.027 with SVM using 91,756 strain-specific markers features over 20 independent runs of 10-fold cross-validations, where cross-validations gave much more optimistic results, and many more features were adopted. The experiments show that *group-specific 40*-mers achieved better classification performance with fewer features.

Group-Specific Sequences

The identified *group-specific 40*-mers were assembled into *group-specific* sequences using CAP3 (Huang and Madan, 1999), in which 11 assembled sequences were longer than 200 bp, with length from 210 to 350 bp (available in **Supplementary File S2**). They were aligned by the sequencing reads from the training and validation sets and the independent testing sets. The coverage distributions over the 11 sequences across all samples were represented as heatmaps in **Figure 4**. A noticeable difference appears between the two groups. In the group of healthy individuals, the reads of most samples cannot be aligned to the 11 sequences. In the patient group, the 11 sequences were aligned successively by the reads from most patients. The *de novo* and reference-free assembly produces longer *group-specific* sequences, which enables the discovery of biomarkers.

Taxonomic Information of the *Group-Specific* Markers

We aligned the 11 *LC-specific* sequences to genomes with "Nucleotide Blast" in NCBI, and all of the sequences were aligned to two strains of *V. parvula*, UTDB1-3, and DSM2008, with 100% query coverage and 97–100% identity. In a previous analysis based on the alignments from reads to reference genomes (Qin et al., 2014), *V. parvula* demonstrated a significant difference in abundance between the two groups of LC patients and healthy individuals.

All 37,302 *group-specific* logical features and 345 *group-specific* numerical features were also blasted to reference genomes in NCBI, 31,067 of logical and 268 of numerical *40*-mers could be matched to *V. parvula* strain UTDB1-3, and 29,712 of logical and 262 of numerical *40*-mers could be matched to *V. parvula* strain DSM2008. Using *V. parvula* strain UTDB1-3 as an example, **Figure 5A** shows the coverage of the whole genome (2.17 Mbp) by the *LC-specific 40*-mers. The horizontal axis is the whole genome. The *40*-mers covered most parts of the genome. **Figures 5B–D** are the zoomed-in alignments and coverages of the genome: 108,308–122,356, 2,037,894–2,038,165, and 2,038,052–2,038,119, marked as "zoom1," "zoom2," and "zoom3", respectively, in the figure. It is clear that many regions are highly and consecutively covered by *k*-mers. As shown in **Figure 5E**, region 1,423,893–1,423,993 of *V. parvula* strain DSM2008 corresponds to "Zoom3" region of *V. parvula* strain UTDB1-3. Comparing the regions in these two strains, the consensus mismatch against UTDB1 is absent on DSM2008, while DSM2008 presents another consensus mismatch against

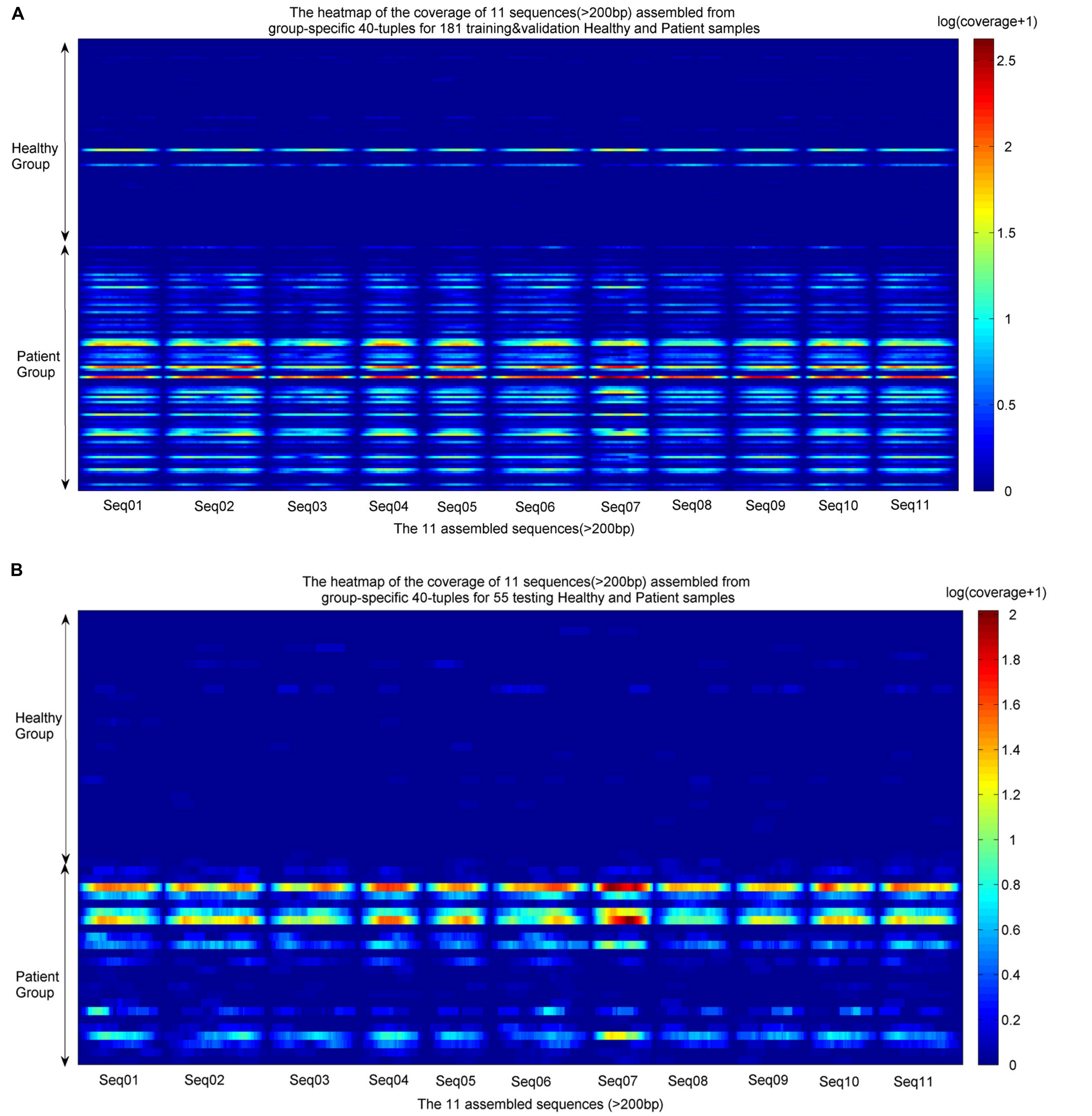


FIGURE 4 | Heatmaps of coverage distribution over the 11 assembled sequences by the metagenomic reads from the training, validation, and testing samples. **(A)** Heatmap of the reads coverage of the 11 assembled sequences across the training and validation samples (83 healthy individuals and 98 LC patients). **(B)** Heatmap of the reads coverage of the 11 assembled sequences across the testing samples (30 healthy individuals and 25 LC patients). The coverage is the read-alignment depth in each nucleotide normalized by the number of million reads. To avoid the effect of large span, we use the logarithm of (coverage+1) as the numerical value of the heatmaps. The horizontal axis is composed of each nucleotide of the 11 sequences, and the vertical axis is composed of healthy individuals and patients. The upper part of each heatmap is the healthy group, and the lower part is the patient group.

DSM2008: 1,423,924. The consistent mismatches against strains UTDB1 and DSM2008 in *V. parvula* indicate the possible existence of an unknown strain of *V. parvula*, which would exist in the gut of LC patients but be absent in the gut of healthy controls.

The IBD-Associated and WT2D-Associated Metagenomic Datasets

The additional two disease-associated metagenomic datasets were analyzed with 20 independent runs of 10-fold cross-validation to evaluate the classification performance for easy

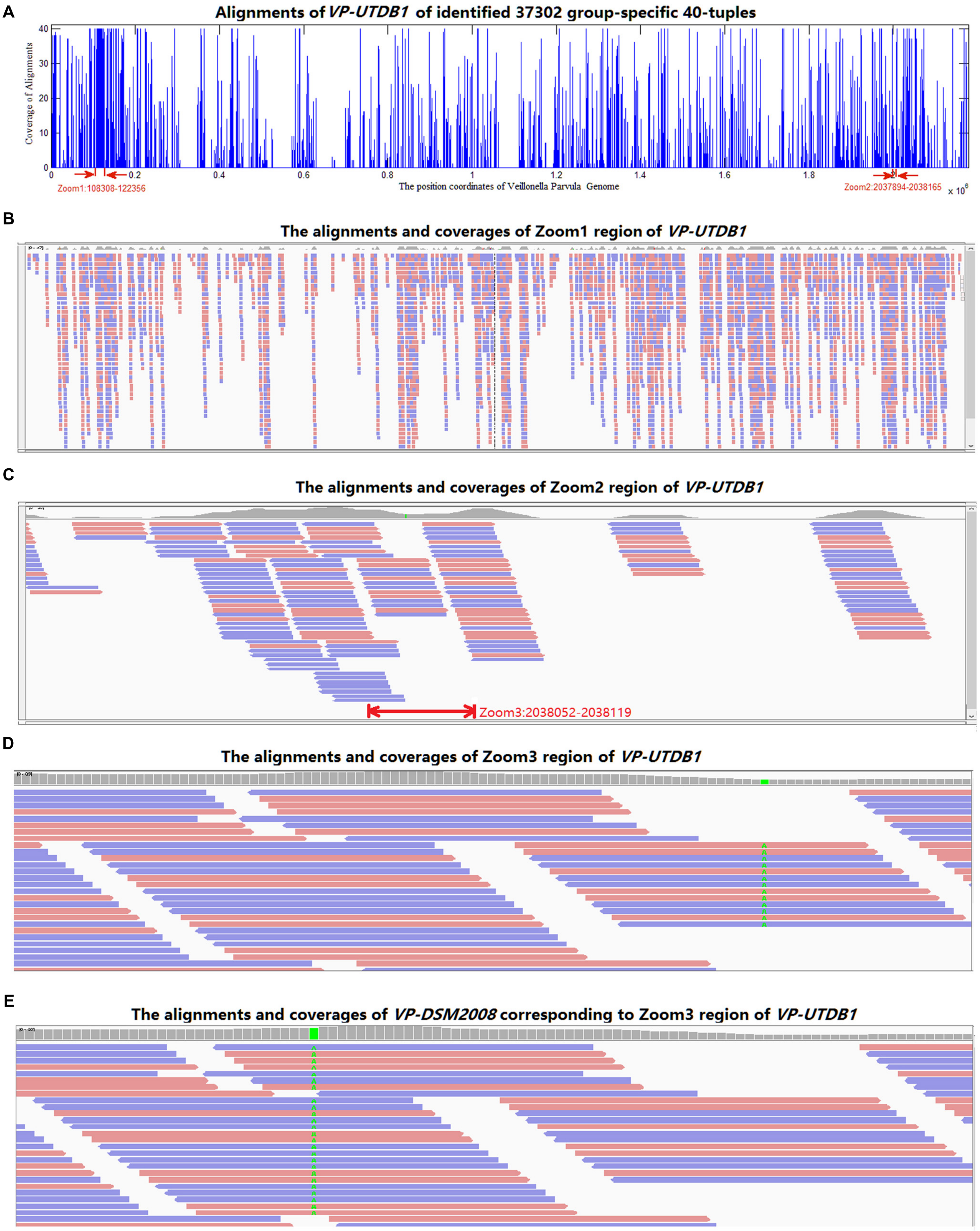


FIGURE 5 | The alignments of the identified *group-specific 40*-mers to the genome sequence of *V. parvula* strain UTDB1-3. **(A)** The alignment distribution over the whole genome. **(B)** The alignments and coverages of region 108,308–122,356 (Zoom1). The red and blue bars denote the *40*-mers matched to reference genome sequence forward and backward, respectively. **(C)** The alignments and coverages of region 2,037,894–2,038,165 (Zoom2). **(D)** The alignments and coverages of region 2,038,053–2,038,119 (Zoom3) with consensus mismatches on 2,038,082. **(E)** The alignments and coverages of region 1,423,893–1,423,993 of *V. parvula* strain DSM2008. This region corresponds to the Zoom3 region of *V. parvula* strain UTDB1-3. Comparing the two regions in the two strains, the consensus mismatch (in green color in **D**) on UTDB1 is absent on DSM2008, but DSM2008 presents another consensus mismatch (in green color in **E**) on DSM2008: 1,423,924.

comparison with previous studies. We emphasized that feature preprocessing and selection were done using only the training set, thereby avoiding biased and overly optimistic performance (Zhang et al., 2006; Pasolli et al., 2016).

The IBD-Associated Dataset

For each fold test of 10-fold cross-validation, about 7000 *group-specific* logical features with ASS ≥ 0.8, but no *group-specific* numerical features, were identified. The numbers of *group-specific* features varied with different fold tests. Because of the relatively small sample size, *30*-mers were set as features. For each *group-specific 30*-mer, its *single-logical-feature* predictor yielded an ASS score on validation. For each round of cross-validation, ∼7000×10 (∼7000 *single-logical-feature* predictors and 10-folds) ASS values were obtained on validations. The boxplots in **Figure 6A** present the distribution of the ∼70,000 ASS values in 20 rounds of 10-fold cross-validation. The values are between 0.78 and 0.89, and they centered at 0.81–0.82, indicating that individual binary features can achieve ASS ≥ 0.78 solely on validation. The average ASS score is 0.875 ± 0.004 (95% confidence interval). The top 15 ranked features were combined to design a random forests classifier. **Figure 6B** presents the ROC curves of 20 independent runs, which were averaged over the 10-folds of cross-validation. The mean AUC of 20 runs is 0.990 ± 0.005 (95% confidence interval), which is much higher than the results reported in previous studies. As shown in **Table 4**, using the same dataset, Pasolli et al. (2016) designed two classifiers. The random forests classifier based on 443 species-abundance features achieved an averaged AUC = 0.893 ± 0.080 under the same experimental setting. The SVM classifier based on the presence of 91,756 strain-specific markers achieved AUC = 0.914 ± 0.084. Xing et al. (2017) obtained AUC = 0.967 with a logistic regression model with LASSO penalty in leave-one-out cross-validation (LOOCV), which used the relative abundances of bins as features. In another study, Cui and Zhang (2013) obtained accuracy = 88%, sensitivity = 92%, and specificity = 84% with 200 *7*-mer features at LOOCV on 25 healthy subjects and 25 patients, where the samples were the subset of our experiment and LOOCV was more relaxed than 10-fold cross-validation.

The WT2D-Associated Dataset

For each fold test of 10-fold cross-validation, ∼700 *40*-mers with ASS ≥ 0.75 were identified, and the best ASS score was 0.78. The classifier designed with random forests using 10 top *group-specific 40*-mer features obtained an average AUC = 0.939 ± 0.011 on the 20 independent runs of 10-fold cross-validation, as shown in **Figure 6C**. In previous studies under the same experimental setting, the average AUCs were 0.834 using 50 metagenomic clusters as features (Karlsson et al., 2013) and 0.785 ± 0.104 using the presence of 83,456 strain-specific markers as features (Pasolli et al., 2016). For further comparison, we implemented metagenome-wide *de novo* assembly with MegaHIT (Li et al., 2015) and then binned the contigs with MetaGen (Xing et al., 2017). The relative abundances of bins were used as features to separate the patient and control groups. The total of 96 samples were too large for read assembly, which required >256 GB memory for 80 samples, and the alignments of reads to the contigs were time-consuming. Therefore, 20 patients and 20 healthy individuals were randomly selected as the training set. The remaining 56 samples were used for independent testing. The relative abundances of bins generated by MetaGen were used as features and the random forests classifier was designed on the training set. The definition of relative abundance in MetaGen includes the parameters that should be determined for each species (they assumed each bin is each species) and each sample through the algorithm of MetaGen. When the classifier was tested on the independent set, these parameters for independent samples are also required to be determined. Personal communications with MetaGen's developers, we revised the code of MetaGen and calculated the feature values of the relative abundances of selected bins for each testing sample. With random forests, MetaGen achieved AUC = 0.685 using 3 features of bins and AUC = 0.735 using 15 features of bins on testing data. With the same training samples, our pipeline obtained AUC = 0.782 with 3 features of *k*-mers and AUC = 0.794 using 15 features of *k*-mers with random forests on testing data. Although both methods are reference free, the *group-specific k*-mers show greater discriminative power than the contig bins for predicting the disease status. Besides, the *de novo* assembly and contig binning are time-consuming. For example, it took 120 h to finish the running from read assembly to contig binning on this training set.

From the experiments, IBD is more predictable than T2D. The experiments on the two disease-associated datasets demonstrate that *group-specific k*-mers achieved much better classification performance with fewer features than previous studies that used the features of short *k*-mer frequencies, species abundance, and strain marker presence. The experiments confirm the effectiveness of long *k*-mer features and the strategy of identifying *group-specific* features.

Running the Computational Pipeline on *Apache Spark*

For the LC dataset, it took 65 h to identify the *group-specific 40*-mers from 56 healthy and 66 LC training samples (252 GB fasta.gz files), including the calculation of *40*-mer frequency vector, the integration of feature matrix, and the identification of the *group-specific 40*-mers. The peak storage space is about 1.5 TB. The above result was run on a local mode of a server with 128 G-memory and Intel(R) Xeon(R) CPU E5-2620 v4 with 8 CPU cores at 2.10 GHz.

DISCUSSION

Different diseases have different levels of association-complexity with human microbiome. If one disease is significantly associated with a specific microbial strain/species/gene, then the disease is highly predictable using a *single-feature* predictor. That is, the disease can be diagnosed with a single microbial biomarker. However, many human diseases are complex in the sense that multiple *group-specific* markers are required to characterize the relevance of disease and microbiome. For these diseases, we have

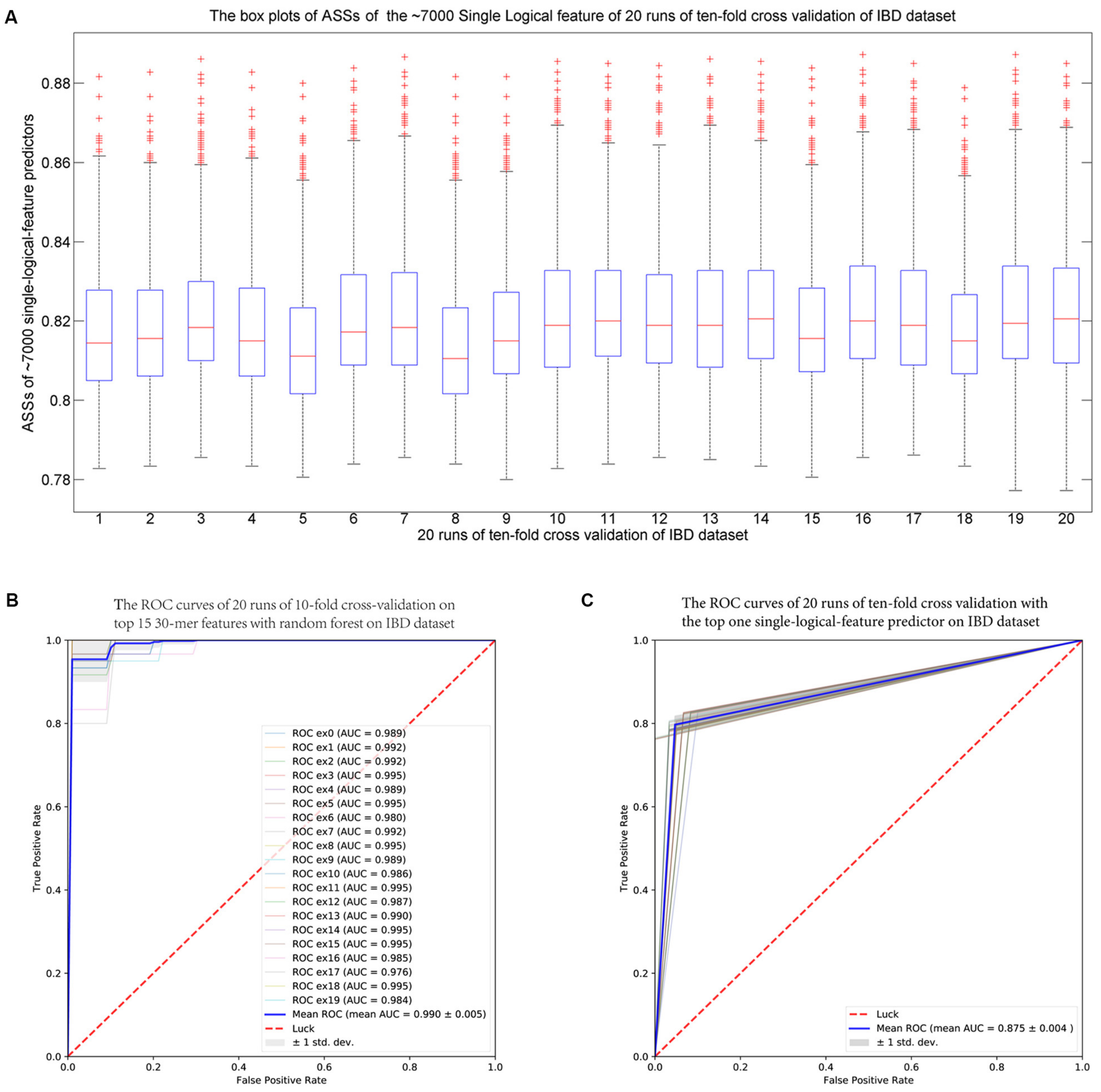


FIGURE 6 | (A) The IBD-associated dataset: the boxplots of ASS by *single-logical-feature* predictors on each one of the identified ~7000 *group-specific* features in the 20 independent runs of 10-fold cross-validation on the IBD dataset. Each boxplot is composed of ~70,000 ASS values on each round of cross-validation. The ASS values are between 0.78 and 0.89 and centered on 0.81–0.82. The "+" symbol denotes outliers. **(B)** The ROC curves of the IBD-associated dataset: The top 15 ranked *30*-mers were combined to design the random forests classifier. The 20 ROC curves are from the 20 independent runs, and each one is the average over the 10-folds of cross-validation. The mean AUC is 0.990 ± 0.005 (95% confidence interval). **(C)** The ROC curves of the WT2D-associated dataset: The top 10 ranked *40*-mers were combined to design the random forests classifier. The 20 ROC curves are from the 20 independent runs, and each one is the average over the 10-folds of cross-validation. The mean AUC is 0.939 ± 0.011 (95% confidence interval).

shown that combining several *group-specific* features can improve prediction accuracy.

In MetaGO, features were selected based on three preset thresholds, including ASS of *single-logical-feature* predictor (θ_1), *p*-value of Wilcoxon rank-sum test for numerical features (θ_2), and *single-numerical* logistic-regression predictor (θ_3). For the IBD-associated and LC-associated datasets, we set $\theta_1 = 0.8$, $\theta_2 = 0.01$, and $\theta_3 = 0.8$, respectively. However, for diseases having more complex associations with microbiome, such as T2D (Pasolli et al., 2016), θ_1 was relaxed to 0.75, $\theta_2 = 0.05$ and $\theta_3 = 0.75$. Therefore, the three thresholds were, in effect, set according to the expected discriminant power of features and the complexity of association between disease and microbiome.

TABLE 4 | Comparison of performance of different methods based on the IBD and WT2D datasets.

	IBD dataset					
Experiment	**20 runs of 10-fold cross-validation (25P+97H)**					**Five runs of LOOCV (25P+25H)**
Feature	***30*-mer**		***30*-mer**	Species abundance†	Presence of strain-specific markers†	Abundance in contig bin††††	*7*-mer††
Number of feature	**1**		**15**	443	91756	Not mentioned	200
Classifier	**Single logical feature predictor**		**Random forests**	Random forests	Support vector machine	Logistic regression + LASSO	Support vector machine
AUC	**ASS* = 0.875 ± 0.004**		**0.990 ± 0.005**	0.893 ± 0.080	0.914 ± 0.084	0.967	Accuracy = 0.88
	WT2D dataset						
Experiment	**20 runs of 10-fold cross-validation (52P+43H)**					**Training (20H+20P)Testing (32P+13H)**	
Feature	***40*-mer**	***40*-mer**	Species abundance†	Presence of strain-specific markers†	Gene markers†††	Abundance of bins with MetaGen	***40-mer***
Number of feature	**1**	**10**	381	83456	50	3	**3**
Classifier	**Single logical feature predictor**	**Random forests**	Random forests	Support vector machine	Support vector machine	Random forests	**Random forests**
AUC	**ASS = 0.76 ± 0.003**	**0.939 ± 0.011**	0.772 ± 0.116	0.785 ± 0.104	0.83	0.961 (training) 0.685 (testing)	**0.979 (training)** **0.782 (testing)**

Using much fewer features, MetaGO achieved better results compared to other methods. The results of MetaGO were in bold. There were two experimental setting for IBD dataset, the "Five runs of LOOCV" are the subset of our experiment and LOOCV was more relaxed than 10-fold cross-validation. For the WT2D dataset, 40-mers were tested under two experimental setting for comparing with other methods. †*(Pasolli et al., 2016);* ††*(Cui and Zhang, 2013);* †††*(Qin et al., 2014);* ††††*(Xing et al., 2017); *average of sensitivity and specificity.*

MetaGO was designed and implemented for two-group case and control datasets. For some studies, there may exist multiple subgroups for the disease, or a pre-disease group. An example of subgroups for disease is the AR-type (marked akinesia and rigidity) and T-type (predominant resting tremor) in Parkinson's disease (Paulus and Jellinger, 1991). Two examples of pre-disease state are impaired glucose tolerance state between T2D and normal glucose tolerance (Karlsson et al., 2013) and colorectal adenoma state between carcinoma and healthy state (Feng et al., 2015). For the multiple-groups scenario, the way to use MetaGO depends on the research purpose. If the purpose is to identify some microbial organisms that are associated with all sub-groups of the disease, we can combine all individuals belonging to any disease groups and treat them as one disease group. MetaGO can be used to the disease and control groups to identify the common microbial organisms associated with all groups of diseases. On the other hand, if the purpose is to identify certain microbial organisms that are specific to a particular group, we can combine all other individuals into one group and then use MetaGO to identify group-specific-associated microbial organisms. Extending MetaGO for a joint analysis of *group-specific* organisms in all the control and different disease groups is a topic of further study.

CONCLUSION

In this study, we developed a computational framework, MetaGO, that is free from reference sequences, metagenome-wide *de novo* assembly, and sequence alignment, to identify *group-specific* sequences between two groups of microbial communities using long *k*-mer features. The *k*-mer length was set between 30 and 40 based on the tradeoff among sensitivity, specificity, and computational cost. The identified *group-specific k*-mers present improved discriminant power for diagnosing diseases using human gut metagenomics data compared with previous studies.

To overcome the computational challenge of long *k*-mer features, an open-source, parallel-computing pipeline was developed on *Apache Spark* to save computational resources and reduce running time. In this study, we applied MetaGO to analyze metagenomic disease-associated datasets. It should be noted that the pipeline is also suitable for identifying *group-specific k*-mers for all types of high-throughput sequencing data where samples are collected from different groups, such as disease-associated human genome sequencing data or other phenotype-associated metagenomic datasets from different environments.

Our experiments validated improvements made by the identified *group-specific k*-mer features compared to previous studies using other types of features. The *group-specific* sequences offer deep and detailed insights required to understand the differences between groups because the method essentially identifies a sequence that is present, or rich, in one group, but absent, or scarce, in another group, the fundamental working principle of *group-specific* sequences. We found that biological explorations based on *group-specific* sequences are consistent with those from previous biological experiments, but additionally offered the potential for new discoveries. Therefore, using long *k*-mer sequence signatures is an effective way to discover biological features, paving the way for a new paradigm of biomarker discovery in the context of host phenotypes. MetaGO enables the detection of *group-specific* features and development of prediction models using a single feature, or a combination of a few features, which helps to reduce the complexity of the model, while increasing the potential feasibility of follow-up discovery of discriminative microbial biomarker(s) for the easy diagnosis of human diseases.

AVAILABILITY OF SUPPORTING DATA AND SOURCE CODES

Source codes and testing data are available at https://github.com/VVsmileyx/MetaGO. The metagenomic sequencing datasets of IBD, LC, and T2D of European women were from the European Bioinformatics Institute's European Nucleotide Archive under accession numbers (EMBL: ERP000108, ERP005860, and ERP002469).

AUTHOR CONTRIBUTIONS

YW, FS, and TC planned the project. YW and ZY designed the model and experiments. LF performed the experiments. YW, JR, and FS analyzed the data. LF contributed materials/analysis tools. YW, JR, ZY, and FS wrote the main manuscript. All authors read and approved the final manuscript.

FUNDING

The research was supported by the National Natural Science Foundation of China (61673324, 61503314, and 61561146396); U.S. National Institutes of Health R01GM120624 and National Science Foundation DMS-1518001; the Natural Science Foundation of Fujian (2016J01316), and the scholarship from China Scholarship Council (201606315011).

ACKNOWLEDGMENTS

We thank Prof. Fan Yang and Jiping Tao at Xiamen University, China, for helpful discussions and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00872/full#supplementary-material

FILE S1 | Detailed descriptions of method and results.

FILE S2 | *LC-specific 40*-mers and sequences.

REFERENCES

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

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., et al. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput. Sci.* 2:e94. doi: 10.7717/peerj-cs.94

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

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694–1697. doi: 10.1126/science.1177486

Cui, H., and Zhang, X. (2013). Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genomics* 14:641. doi: 10.1186/1471-2164-14-641

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/ncomms7528

Fofanov, Y., Luo, Y., Katili, C., Wang, J., Belosludtsev, Y., Powdrill, T., et al. (2004). How independent are the appearances of n-mers in different genomes? *Bioinformatics* 20, 2421–2428. doi: 10.1093/bioinformatics/bth266

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Han, W., Wang, M., and Ye, Y. (2017). "A concurrent subtractive assembly approach for identification of disease associated sub-metagenomes," in *Research in Computational Molecular Biology. RECOMB 2017. Lecture Notes in Computer Science*, Vol. 10229, ed. S. Sahinalp (Cham: Springer).

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F., and Zhang, X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 13:730. doi: 10.1186/1471-2164-13-730

Jiang, R. (2015). Walking on multiple disease-gene networks to prioritize candidate genes. *J. Mol. Cell Biol.* 7, 214–230. doi: 10.1093/jmcb/mjv008

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

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578. doi: 10.1128/MMBR.00009-08

Le, V. V., Lang, T. V., Le, T. B., and Hoai, T. V. (2015). A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms Mol. Biol.* 10:2. doi: 10.1186/s13015-014-0030-4

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi: 10.1101/gr.097261.109

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

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., et al. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.* 23, 1704–1714. doi: 10.1101/gr.151803.112

Lu, Y. Y., Chen, T., Fuhrman, J. A., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* 33, 791–798. doi: 10.1093/bioinformatics/btw290

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Papudeshi, B., Haggerty, J. M., Doane, M., Morris, M. M., Walsh, K., Beattie, D. T., et al. (2017). Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics* 18:915. doi: 10.1186/s12864-017-4294-1

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14, 1023–1024. doi: 10.1038/nmeth.4468

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

Paulus, W., and Jellinger, K. (1991). The neuropathologic basis of different clinical subgroups of Parkinson's disease. *J. Neuropathol. Exp. Neurol.* 50, 743–755. doi: 10.1097/00005072-199111000-00006

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

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69. doi: 10.1186/s40168-017-0283-5

Richter, D. C., Ott, F., Auch, A. F., Schmid, R., and Huson, D. H. (2008). MetaSim—a sequencing simulator for genomics and metagenomics. *PLoS One* 3:e3373. doi: 10.1371/journal.pone.0003373

Rizk, G., Lavenier, D., and Chikhi, R. (2013). DSK: k-mer counting with very low memory usage. *Bioinformatics* 29, 652–653. doi: 10.1093/bioinformatics/btt020

Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8. doi: 10.1186/s40168-016-0154-5

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

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Wang, Y., Lei, X., Wang, S., Wang, Z., Song, N., Zeng, F., et al. (2015). Effect of k-tuple length on sample-comparison with high-throughput sequencing data. *Biochem. Biophys. Res. Commun.* 469, 1021–1027. doi: 10.1016/j.bbrc.2015.11.094

Wang, Y., Liu, L., Chen, L., Chen, T., and Sun, F. (2014). Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS One* 9:e84348. doi: 10.1371/journal.pone.0084348

Wang, Y., Wang, K., Lu, Y. Y., and Sun, F. (2017). Improving contig binning of metagenomic data using dS2oligonucleotide frequency dissimilarity. *BMC Bioinformatics* 18:425. doi: 10.1186/s12859-017-1835-1

Wen, C., Zheng, Z., Shao, T., Lin, L., Xie, Z., Chatelier, E. L., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* 18:142. doi: 10.1186/s13059-017-1271-6

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wiest, R., Lawson, M., and Geuking, M. (2014). Pathological bacterial translocation in liver cirrhosis. *J. Hepatol.* 60, 197–209. doi: 10.1016/j.jhep.2013.07.044

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Xing, X., Liu, J. S., and Zhong, W. (2017). MetaGen: reference-free learning with multiple metagenomic samples. *Genome Biol.* 18:187. doi: 10.1186/s13059-017-1323-y

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

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: cluster computing with working sets. *HotCloud* 10:95.

Zhang, X., Lu, X., Shi, Q., Xu, X. Q., Leung, H. C., Harris, L. N., et al. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 7:197. doi: 10.1186/1471-2105-7-197

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.

METHODS
published: 01 March 2018
doi: 10.3389/fmicb.2018.00365

BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data

Colin P. McNally[1]†, Alexander Eng[1]†, Cecilia Noecker[1]†, William C. Gagne-Maynard[2] and Elhanan Borenstein[1,3,4]*

[1] Department of Genome Sciences, University of Washington, Seattle, WA, United States, [2] Institute for Health Metrics and Evaluation, Seattle, WA, United States, [3] Department of Computer Science and Engineering, University of Washington, Seattle, WA, United States, [4] Santa Fe Institute, Santa Fe, NM, United States

The abundance of both taxonomic groups and gene categories in microbiome samples can now be easily assayed via various sequencing technologies, and visualized using a variety of software tools. However, the assemblage of taxa in the microbiome and its gene content are clearly linked, and tools for visualizing the relationship between these two facets of microbiome composition and for facilitating exploratory analysis of their co-variation are lacking. Here we introduce *BURRITO*, a web tool for interactive visualization of microbiome multi-omic data with paired taxonomic and functional information. BURRITO simultaneously visualizes the taxonomic and functional compositions of multiple samples and dynamically highlights relationships between taxa and functions to capture the underlying structure of these data. Users can browse for taxa and functions of interest and interactively explore the share of each function attributed to each taxon across samples. BURRITO supports multiple input formats for taxonomic and metagenomic data, allows adjustment of data granularity, and can export generated visualizations as static publication-ready formatted figures. In this paper, we describe the functionality of BURRITO, and provide illustrative examples of its utility for visualizing various trends in the relationship between the composition of taxa and functions in complex microbiomes.

Keywords: microbiome, metagenomics, data visualization, taxonomy, function, web interface

***Correspondence:**
Elhanan Borenstein
elbo@uw.edu

†These authors have contributed equally to this work.

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

Received: *17 October 2017*
Accepted: *15 February 2018*
Published: *01 March 2018*

Citation:
McNally CP, Eng A, Noecker C, Gagne-Maynard WC and Borenstein E (2018) BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data. Front. Microbiol. 9:365. doi: 10.3389/fmicb.2018.00365

BACKGROUND

Microbial communities are complex ecosystems with important impacts on human health and on the environment. High-throughput DNA sequencing has enabled comprehensive profiling of these communities in terms of their composition and structure. Traditionally, microbial ecology studies resort to one of two primary approaches for profiling the composition of a given community, focusing either on its taxonomic composition (e.g., using targeted 16S rRNA gene sequencing or a marker-gene based approach) or on its functional composition (e.g., using metagenomic shotgun sequencing and assessing the abundance of various gene families) (**Figure 1A**). Obtained taxonomic or functional profiles are then often visualized as simple stacked bar or area plots of relative abundances, or via specialized data visualization tools designed for exploring these data.

Abbreviations: BURRITO, Browser Utility for Relating micRobiome Information on Taxonomy and functiOn; KEGG, Kyoto Encyclopedia of Genes and Genomes; OTU, operational taxonomic unit.

For example, Explicet (Robertson et al., 2013) and Krona (Ondov et al., 2011) aid analysis by displaying abundance data while simultaneously presenting the hierarchical relationships between entities. Other tools have gone beyond relative abundance visualization to enable exploration of specific, comparative aspects of microbiome data. EMPeror (Vázquez-Baeza et al., 2013), for example, allows users to generate 3D principal coordinate analysis plots to visualize clustering of, or variation in, taxonomic compositions coupled with trends in any associated metadata. Other tools, including Community-Analyzer (Kuntal et al., 2013) and MetaCoMET (Wang et al., 2016), provide access to multiple types of visualizations of the same microbiome data, each highlighting different aspects of community structure or between-sample relationships.

Importantly, however, recent years have witnessed an explosion of microbiome *multi-omic* studies that aim to describe simultaneously multiple aspects of community structure, including specifically both taxonomic and functional compositions (Huttenhower et al., 2012; Greenblum et al., 2015; Taxis et al., 2015; Pedersen et al., 2016; Zhernakova et al., 2016; Lloyd-Price et al., 2017). Moreover, recently developed methods can now determine both the taxonomic and functional profile of a given community from the same sequencing data, for example, by assigning shotgun metagenomic reads both taxonomic and functional annotations (Abubucker et al., 2012). Notably, these two facets of microbiome composition are not independent since the set of genes found in a metagenome and their abundances is a direct result of the set of genes (and their copy number) encoded by each community member and the relative abundance of each member in the community (**Figure 1B**). Put differently, the abundance of each gene family (or 'function') in the metagenome can be deconvolved into taxon-specific functional profiles in which shares of the gene family's total abundance are attributed to specific taxa of origin (Carr et al., 2013) (**Figure 1C**). This link between taxonomic and functional compositions can be used, for example, to predict functional abundances from 16S rRNA-based taxonomic profiles (Langille et al., 2013) or to identify taxonomic drivers of disease-associated functional shifts (Manor and Borenstein, 2017b). Importantly, this inherent relationship between taxonomic and functional profiles must also be considered when exploring how functional capacity co-varies with taxonomic composition across samples, since differences in gene abundances between samples are mainly derived from differences in taxonomic composition (Turnbaugh et al., 2009; Oh et al., 2014; Bäckhed et al., 2015). Yet, despite the growing appreciation for this link between taxonomic and functional compositions, an integrative tool that can simultaneously visualize both taxonomic and functional data and that can *account for* and *expose* the relationships between taxonomic and functional variation is lacking.

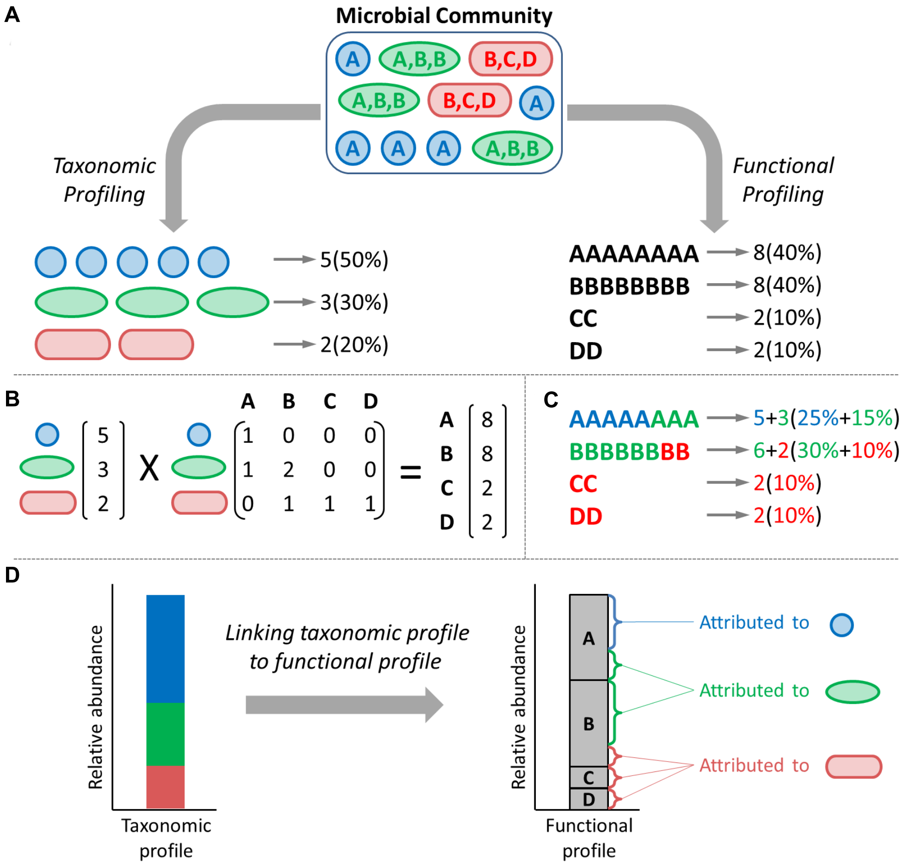
Here we introduce BURRITO, a web-based visualization tool that enables easy and intuitive exploratory analysis of the relationships between taxonomic and functional abundances across microbiome samples. BURRITO simultaneously provides a traditional interface for exploring taxonomic and functional abundances independently while also visualizing the links between these two microbiome facets and highlighting the share of each function's total abundance that is attributed to each taxon (**Figure 1D**). Through an interactive interface, BURRITO also provides ample and precise information about such *attributions* (i.e., the share of each function's total abundance attributed to each taxon), as well as various summary statistics. To facilitate interactive data exploration and publication-quality figure generation for a wide audience, BURRITO further offers multiple options for data input and supports customizing various aspects of the visualization.

METHODS AND IMPLEMENTATION

User Input and Taxa–Function Mapping

BURRITO accommodates multiple types of input data and, depending on the provided data, uses different approaches to attribute the provided or inferred function abundances to taxa of origin (see **Figure 1D**). Specifically, the user can select one of three options for input data and for determining taxa–function attributions (**Figure 2**). In the first option, which requires the bare minimum in terms of input data, the user can simply provide a table of taxonomic abundances across samples (measured as either absolute read counts or relative abundances) using Greengenes 97% OTU IDs for each taxon (DeSantis et al., 2006). Given these taxonomic data, BURRITO will automatically predict the functional profile of each sample and will determine each function's taxonomic attributions using a database of pre-annotated genomic content (following the approach described in **Figures 1B,C**). Briefly, in this approach, taxonomic abundances for each sample are first corrected by dividing each taxon's abundance by its estimated 16S rRNA copy number. The abundance of a given function that is attributed to each taxon (and ultimately the total abundance of that function) is then inferred by multiplying the corrected taxonomic abundances by the number of genes associated with that function in each taxon's genomes. The gene content for each taxon is obtained from PICRUSt (Langille et al., 2013) and functional annotations are based on KEGG Orthology groups (Kanehisa and Goto, 2000; Kanehisa et al., 2015).

The second available option is similar to the one described above, but relies on user-provided genomic content annotations (instead of PICRUSt-based inferred content) to calculate functional profiles and the share of the functional profile attributed to each taxon. This approach is appropriate, for example, for exploring communities with members that may not be adequately represented by PICRUSt-inferred genomes but that can be better characterized based on user proprietary data. As in the first option, the user is required to provide a taxonomic abundance table, but also provides a custom genomic content table describing the set of genes (and their copy number) encoded by each taxon. Given these data, functional abundances and attributions are calculated in the same manner as described above. When using this option, it is also assumed that taxonomic abundances are already normalized by estimated 16S rRNA copy number. Moreover, using this approach the user is not limited to Greengenes OTU IDs or to KEGG Orthology



$$\begin{bmatrix}5\\3\\2\end{bmatrix} \times \begin{array}{c}\begin{array}{cccc}A & B & C & D\end{array}\\ \begin{pmatrix}1 & 0 & 0 & 0\\1 & 2 & 0 & 0\\0 & 1 & 1 & 1\end{pmatrix}\end{array} = \begin{array}{c|c} A & 8\\ B & 8\\ C & 2\\ D & 2\end{array}$$

FIGURE 1 | The taxonomic and functional compositions of a microbiome are inherently linked. **(A)** Typical microbiome studies quantify and report the taxonomic (colored shapes) and functional (letters) profiles of a given community as separate entities. **(B)** The functional profile of a community is a linear combination of the taxonomic composition and the genomic content of each taxon. **(C)** Functional profiles can be deconvolved into taxon-specific functional profiles, denoting which share of the abundance of each function is *attributed* to each taxon. **(D)** Such deconvolved functional profiles can be visualized, illustrating the total abundance of each function as a stacked bar of taxon-specific attributions.

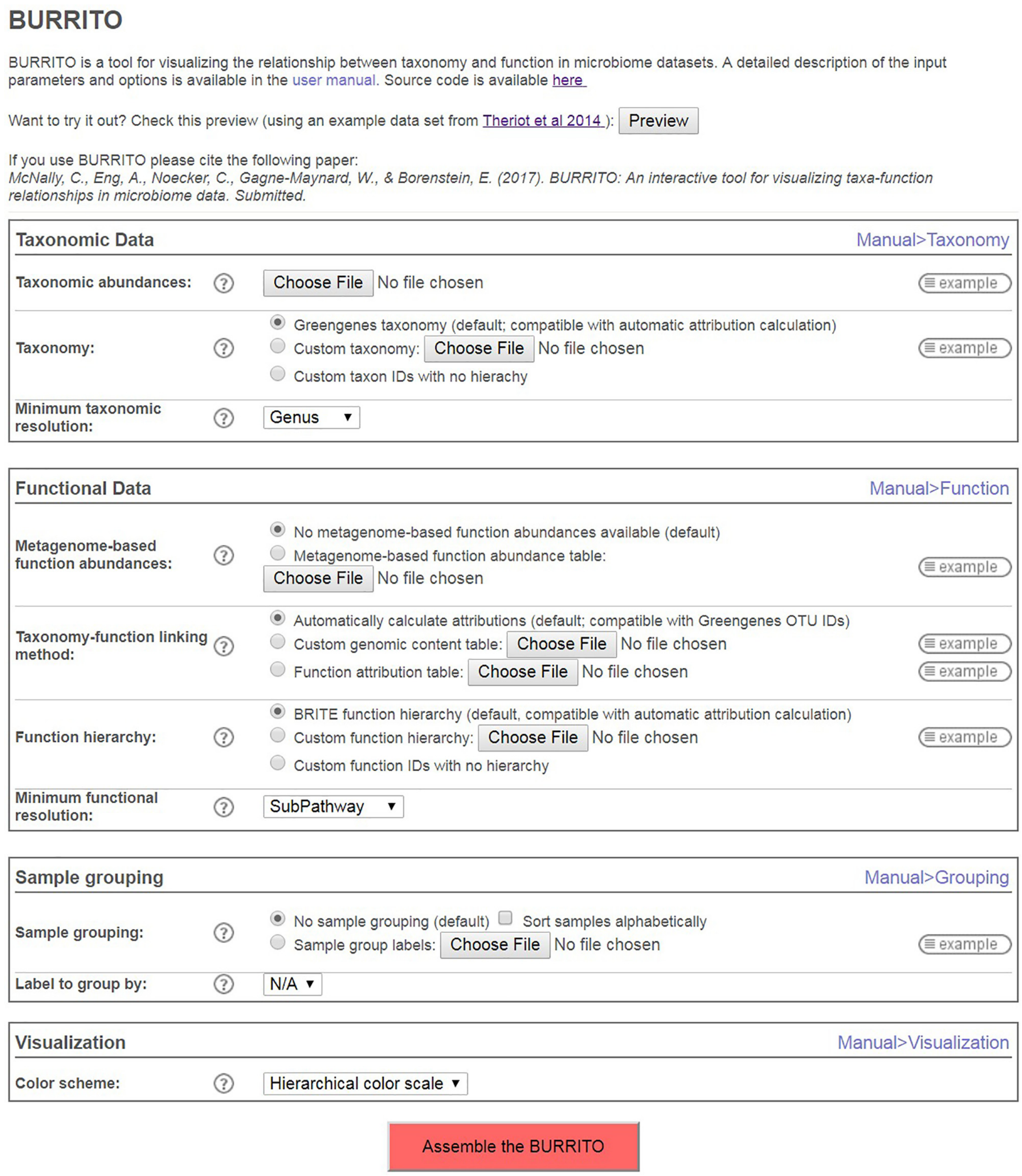
groups, and alternately can use their taxonomic and/or functional classification of choice (as long as the same IDs are used in all relevant input files). Notably, when using custom taxonomic classification or custom hierarchical function relationships, the user can also provide files describing these custom systems to allow taxonomic and functional data to be grouped at different levels (see section "Exploring Attributions at Different Taxonomic and Functional Levels").

The third and final option relies on a pre-determined table of taxon-specific functional attributions rather than calculating attributions from taxonomic composition. This approach may be appropriate, for example, in cases where the user wishes to introduce specific custom modifications to a pre-calculated attribution table. Using this option requires a taxonomic abundance table (as above) and, instead of a genomic content table, a table of taxa-specific functional abundance attributions. As in the second approach, any taxonomic classification or functional hierarchy system can be used.

The specific format for each data input file and the specific restrictions associated with each approach are all noted on BURRITO's upload page and are described in more detail in BURRITO's documentation. Example data files are also available for download from the upload page. Notably, in all three approaches, the user can also upload an independent, paired dataset of functional abundance profiles (i.e., based on functional annotation of shotgun metagenomic sequencing), which will be visualized as described below alongside the calculated taxa-based functional profiles.

Importantly, predicting functional abundances based on taxonomic composition has various drawbacks, including, most notably, limited accuracy that can vary across samples and functions. For example, predicting the functional content of taxa that match available reference genomes (or that are phylogenetically close to sequenced strains) would likely be much more precise than taxa for which only distantly related reference genomes are available. Predicted functional abundance should therefore be considered probabilistic in nature, and indeed tools like PICRUSt provide additional information to describe the confidence of obtained predictions (some of which is made available via BURRITO as described below). With that in mind, BURRITO's option to compare taxa-based functional prediction with functional annotation from metagenomic data could be particularly useful, allowing users to explore prediction accuracy in their data.

In addition to the above primary input, BURRITO's upload page (**Figure 2**) further includes several visualization options, allowing users to better control the way data will be displayed. Specifically, BURRITO supports grouping samples based on user-provided labels (e.g., cases vs. control or conditions). Additionally, the user can select the minimum taxonomic and functional resolution to be displayed. Using a finer resolution allows exploring the data in more depth, but could slow



BURRITO

BURRITO is a tool for visualizing the relationship between taxonomy and function in microbiome datasets. A detailed description of the input parameters and options is available in the user manual. Source code is available here

Want to try it out? Check this preview (using an example data set from Theriot et al 2014): Preview

If you use BURRITO please cite the following paper:
McNally, C., Eng, A., Noecker, C., Gagne-Maynard, W., & Borenstein, E. (2017). BURRITO: An interactive tool for visualizing taxa-function relationships in microbiome data. Submitted.

Taxonomic Data — Manual>Taxonomy

Taxonomic abundances: ? Choose File No file chosen — ≡ example

Taxonomy: ?
● Greengenes taxonomy (default; compatible with automatic attribution calculation)
○ Custom taxonomy: Choose File No file chosen — ≡ example
○ Custom taxon IDs with no hierachy

Minimum taxonomic resolution: ? Genus ▾

Functional Data — Manual>Function

Metagenome-based function abundances: ?
● No metagenome-based function abundances available (default)
○ Metagenome-based function abundance table:
Choose File No file chosen — ≡ example

Taxonomy-function linking method: ?
● Automatically calculate attributions (default; compatible with Greengenes OTU IDs)
○ Custom genomic content table: Choose File No file chosen — ≡ example
○ Function attribution table: Choose File No file chosen — ≡ example

Function hierarchy: ?
● BRITE function hierarchy (default, compatible with automatic attribution calculation)
○ Custom function hierarchy: Choose File No file chosen — ≡ example
○ Custom function IDs with no hierarchy

Minimum functional resolution: ? SubPathway ▾

Sample grouping — Manual>Grouping

Sample grouping: ?
● No sample grouping (default) □ Sort samples alphabetically
○ Sample group labels: Choose File No file chosen — ≡ example

Label to group by: ? N/A ▾

Visualization — Manual>Visualization

Color scheme: ? Hierarchical color scale ▾

Assemble the BURRITO

FIGURE 2 | BURRITO's upload page, describing the various input approaches BURRITO supports and other visualization and analysis options.

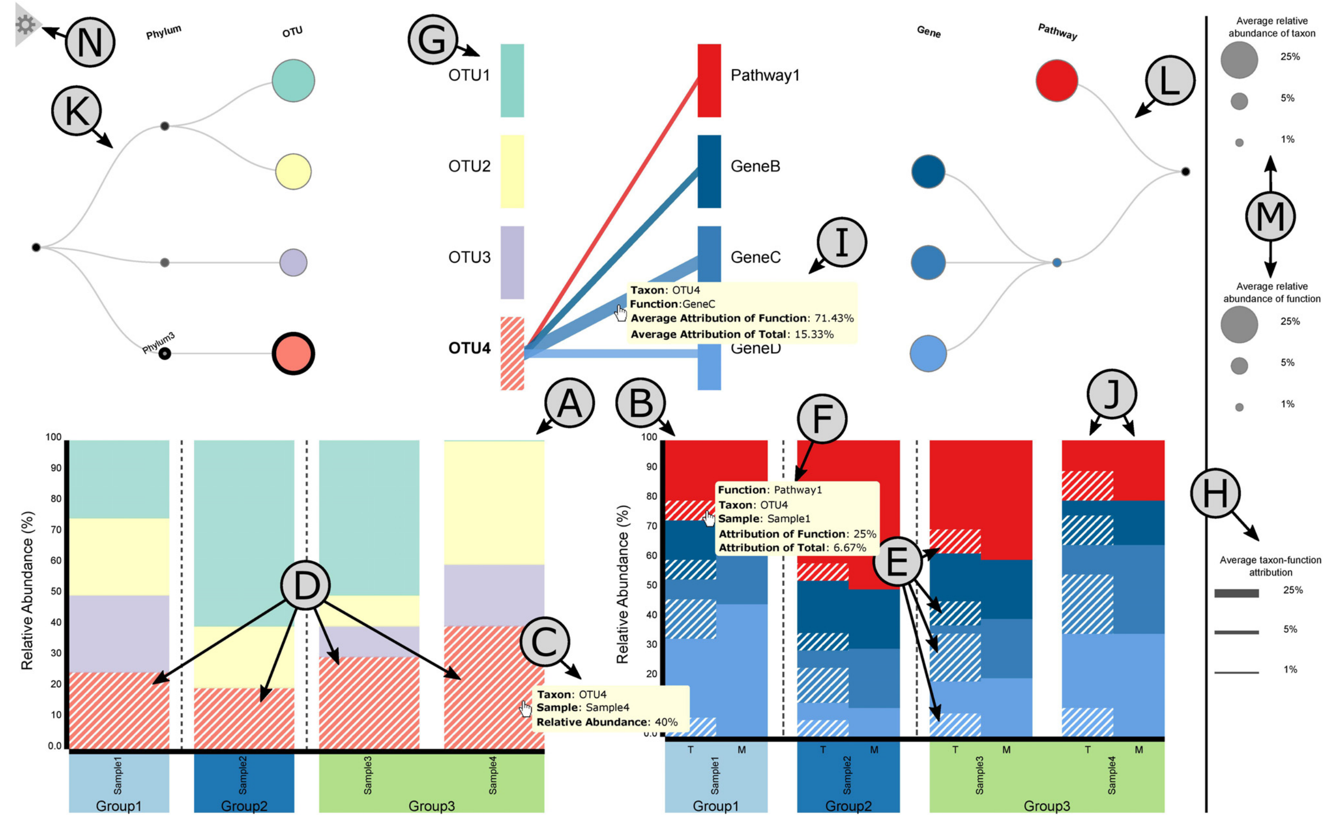
visualization performance due to the large number of elements that need to be calculated, managed, and displayed. Finally, users can choose between hierarchical or random color schemes to distinguish taxa and functions.

Visualizing the Relationship Between Taxonomic and Functional Abundances

BURRITO uses two stacked bar plots, one for taxonomy (**Figure 3A**) and one for function (**Figure 3B**), to provide a standard visualization of taxonomic and functional relative abundances in each sample. Taxonomic abundances are taken from the user-provided taxonomic abundance table. Functional abundances are calculated as described in the previous sections and are displayed as the sums of all taxon-specific attributions for each function and sample. Precise abundance values for each taxon or function in each sample can be viewed in a *tooltip* that appears when the user hovers over any bar segment (**Figure 3C**). Additionally, hovering over a bar segment highlights the corresponding taxon's or function's bar segment across all samples to aid visual comparison of abundances between samples (**Figure 3D**).

The most innovative component of BURRITO is the visualization of how function abundance shares are attributed to the various taxa. This information is revealed when the user hovers over a bar segment in the taxonomic abundance bar plot, which, in addition to highlighting taxon abundances as noted above, also highlights the portion of each function abundance bar segment (in each sample) that is attributed to this taxon (**Figure 3E**). To view the exact function abundance share attributed to a given taxon, the user can click on (rather than hover over) a taxon's bar segment to lock this taxon-specific attribution highlighting, and then hover over the highlighted portion of a function bar segment, revealing a tooltip with the corresponding information (**Figure 3F**).

BURRITO also displays a "control panel" that can be used to investigate specific taxa and/or functions and explore average abundances across samples (**Figure 3G**). Specifically, taxa and functions are represented by bars on the left and right sides



Taxon: OTU4
Function:GeneC
Average Attribution of Function: 71.43%
Average Attribution of Total: 15.33%

Function: Pathway1
Taxon: OTU4
Sample: Sample1
Attribution of Function: 25%
Attribution of Total: 6.67%

Taxon: OTU4
Sample: Sample4
Relative Abundance: 40%

FIGURE 3 | A layout of BURRITO's visualization. **(A,B)** Stacked bar plots of taxonomic and functional composition across samples. **(C)** Tooltips appear when hovering over each taxon, providing information about this taxon's relative abundance in each sample. **(D)** Interactive highlighting of individual taxa, which correspondingly highlights the shares of functional abundances that can be attributed to the taxon in question **(E)**. **(F)** Tooltips provide detailed function abundance and attribution data for each sample. **(G)** The bipartite graph control panel identifies individual taxa and functions and shows links between them. **(H)** Edge width in the control panel represents the *average* share of a function that is attributed to a given taxon. **(I)** Exact taxon-function attribution values can be seen by hovering over the edge connecting the taxon to the function. **(J)** An independent dataset of shotgun metagenomics-derived function abundances can be provided and displayed alongside inferred taxon-specific function abundances. **(K,L)** The data can be viewed at a higher or lower taxonomic or functional resolution by clicking on nodes in the corresponding tree diagrams. **(M)** The size of each node in the tree represents the average abundance of that entity. **(N)** Opening the hidden menu allows users to export visualization plots and processed data.

of a bipartite graph. Hovering over the control panel performs similar highlighting as the bar plots, highlighting the linked abundances of individual taxa and functions and displaying (via a tooltip) the average relative abundance of each taxon or function. Moreover, when a taxon is highlighted, edges that link that taxon to all the functions it encodes are displayed, providing an easy reference for identifying the functions with shares attributed to that taxon. Similarly, when a function is highlighted, edges that indicate which taxa encode that function (and hence have shares of that function attributed to them) are displayed. The width of an edge between a taxon and a function represents the average share of that function's abundance that is attributed to that taxon across all samples (**Figure 3H**). Clicking on a specific taxon or function in the control panel (or bar plots) locks the selection of that taxon or function, allowing the user to hover over each edge and view (via a tooltip) the exact taxon-function attribution values (**Figure 3I**). Additionally, when a taxon or function selection has been clicked (i.e., selected) and edges connecting taxa and functions are displayed, the user can highlight the abundances of a single function attributed to a single taxon by clicking on the edge between them.

BURRITO also supports comparison between taxa-based function abundances (calculated based on taxonomic profiles and genomic content as noted above) and a separate dataset of user-provided function abundances (typically obtained by functional annotation of shotgun metagenomic sequencing reads). If such a dataset is included in the input, these user-provided function abundances are displayed adjacent to the taxa-based function abundances for comparison (**Figure 3J**).

Exploring Attributions at Different Taxonomic and Functional Levels

BURRITO provides an interactive and intuitive interface for exploring abundance and attribution data at varying taxonomic and functional resolutions. Taxonomic classification and hierarchical function relationships (either those used by default or custom systems provided by the user as noted above) are each represented as a tree above the corresponding abundance bar plot (**Figures 3K** and **3L**, respectively). To aid the user in understanding how different taxa or functions are related in the bipartite graph and bar plots, these trees are aligned to the left and right of the bipartite graph with all taxa and functions appearing in the same vertical order across all components of the visualization. These trees also indicate average taxon and function abundances across samples by the size of leaf nodes in the trees (**Figure 3M**). Beyond visualizing hierarchical relationships, these trees also provide a tool for interactive data exploration by allowing the user to expand or collapse different leaves or branches of each tree. Clicking on a leaf node reveals all taxonomic or functional subcategories of that node in the tree, and correspondingly expands the bipartite graph and subdivides the relevant relative abundance bars in the bar plot into the relative abundances of those subcategories. Alternatively, clicking on a non-leaf node within a tree performs the reverse operation, collapsing all visible descendants of that node, making the clicked node a leaf node, and aggregating the abundance bars for those descendants into the abundance bars for the clicked node. Importantly, these interactive features allow the user to dynamically drill up or down in both taxonomic and functional resolution across the different branches of either tree as they explore the data.

Exporting Visualization Plots and Processed Data

BURRITO also provides options for exporting a static version of the visualization (e.g., for including in presentations or publications), for downloading the function abundance table or attribution table underlying the displayed function plot, and for downloading basic statistics. These options can be accessed from the visualization screen via a hidden menu (**Figure 3N**). Exported figures will maintain any currently-selected highlighting and taxonomic or functional tree expansion. In addition to exporting the full visualization, users can choose to individually export either bar plot of relative abundances and either half of the bipartite graph, which can serve as a legend for the color-coding of taxa or functions in the bar plots. All images can be exported in PNG or SVG format.

If users wish to further explore the predicted function abundance or attribution data in more detail, they can also download the tables underlying the visualization. Function abundance and attribution tables can be downloaded at the minimum functional resolution (and minimum taxonomic resolution for the attribution table) specified on the upload page.

Finally, if a binary categorical variable is selected for sample grouping, BURRITO will perform basic differential abundance testing (and multiple hypothesis testing correction) of taxa and functions and provide the results for download. Additionally, for visualizations showing PICRUSt-inferred functional abundances, BURRITO will calculate and provide the average Nearest Sequenced Taxon Index (NSTI; Langille et al., 2013) for each sample, reflecting the overall confidence in predicted functions. This information can be similarly downloaded via the hidden menu.

Technical Implementation

BURRITO's client is a browser page written in HTML and Javascript, utilizing the d3.js library to display data. User-submitted data are uploaded to an R Shiny server for processing (including, specifically, calculation of attributions) and the results are sent back to the browser for visualization. Additional details concerning BURRITO implementation can be found in the Supplementary Text.

CASE STUDIES AND DISCUSSION

To demonstrate the utility of BURRITO, we describe below its application to two microbiome datasets with varying properties.

Case Study 1: Exploring the Effects of Antibiotic Treatment and Recovery on Inferred Functional Composition in the Mouse Cecum

We used BURRITO to visualize a publicly available dataset of 16S rRNA sequencing data, describing cecal samples from mice treated with antibiotics 2 days and 6 weeks after treatment (labeled 'Abx Day 2' and 'Abx Day 42,' respectively), and time-matched controls (labeled 'Control Day 2' and 'Control Day 42,' respectively) (Theriot et al., 2014). Note that BURRITO can visualize grouping even when samples are partitioned into more than two groups (as is the case in this dataset; **Figure 4**), though it does not provide differential abundance statistics in such settings. This study, which focused on associations of the microbiome with metabolomic data and colonization resistance, confirmed significant community perturbations in response to antibiotics. This dataset is also used in BURRITO's *Preview* option on the upload page (see **Figure 2**), allowing users to examine BURRITO's visualization and functionality (and to compare those to the examples provided in this case study) without the need to provide any additional data.

Given this dataset, we used the first input approach described above, allowing BURRITO to predict functional abundances (and taxon-specific attributions) based on the default PICRUSt-derived genomic content table. BURRITO's visualization of taxonomic and functional profiles revealed relatively subtle functional variation despite drastic taxonomic variation across samples, a pattern commonly observed in microbiome studies (Manor and Borenstein, 2017a). Specifically, while Abx Day 2 samples are markedly different from, for example, Control Day 2 samples (with the former being dominated by species from the class *Bacilli* and the latter by species from the classes *Clostridia* and *Bacteroidia*), their predicted functional profiles are relatively similar (**Figure 4A**). Hovering over the various taxa in the taxonomic profiles highlighted the shares of the functional profile in the various sample groups that are attributed to microbes from these three classes (*Bacilli*, *Clostridia,* and *Bacteroidia*), demonstrating, for example, that attributions to the *Bacteroidia* species were relatively small compared to their abundance. See, for instance, sample NonAbx29, in which *Bacteroidia* is more abundant than *Clostridia* (54.93 vs. 44.13%), but has a smaller share attributed to it than to *Clostridia* in every functional category (**Figure 4B**). To further illustrate BURRITO's functionality, we then visually searched for functions that are nonetheless differentially abundant between Abx Day 2 samples and other samples, focusing specifically on pathways in the metabolism category. We observed, for example, that Abx Day 2 samples were generally depleted of amino acid metabolism genes (**Figure 4C**). Indeed, a characteristic shift in gut amino acid concentrations has been previously described in response to antibiotic treatment in mice, and since amino acid availability can facilitate infection by enteric pathogens such as *Clostridium difficile*, understanding the taxonomic determinants of this shift is of great interest (Antunes et al., 2011; Jump et al., 2014; Jenior et al., 2017). Selecting this pathway (by clicking on it in the bar plot or in the control panel) and examining the share of this pathway attributed to each taxon (by then clicking on the edge connecting this pathway to each taxon in the control panel), suggested that its depletion in Abx Day 2 samples could be explained by the fact that *Bacilli* contribute less than *Clostridia* to this pathway compared to their abundances. Specifically, we noted that the share of this pathway attributed to *Bacilli* in Abx Day 2 samples is smaller than the share of this pathway attributed to *Clostridia* in Abx Day 42 samples, even though the abundance of *Bacilli* and *Clostridia* in Abx Day 2 and in Abx Day 42 samples, respectively, is comparable (close to 100%) (**Figure 4D**). Similarly, the share of this pathway attributed to *Bacilli* in Abx Day 2 samples is comparable to the share of this pathway attributed to *Clostridia* in Control samples, even though the abundance of *Bacilli* in Abx Day 2 samples is higher than the abundance of *Clostridia* in control samples. This lower proportional contribution indicates fewer genes involved in this pathway in *Bacilli* compared to *Clostridia*. Lastly, we searched for additional functions with higher or lower shares attributed to *Bacilli* by selecting this taxon (again, by hovering over or clicking on it in the control panel or in the bar plot) and examining the width of the attribution edges connecting it to each function. In this setting, it was easy to note that a relatively small share of the cell motility function is attributed to this taxon, compared to, for example, the shares of metabolic functions attributed to it (**Figure 4E**).

Case Study 2: Taxa–Function Relationships in the Human Microbiome Project

We additionally used BURRITO to visualize data from 21 supragingival plaque samples with both 16S rRNA and shotgun metagenomic data downloaded from the Human Microbiome Project (Huttenhower et al., 2012). We first used the 16S rRNA data alone (i.e., again using the first input approach), examining inferred functional profiles and the taxa they are attributed to. As noted above, expanding the taxonomic tree can provide additional details about specific genera to which each function is attributed. Similarly, expanding the functional tree can offer insights into differentially abundant pathways and subpathways. For example, drilling down into subpathways in the Environmental Information Processing category and examining the average share from each subpathway attributed to the various phyla, we noted that the abundance of ABC transporter genes is attributed primarily to Actinobacteria (average attribution across samples is 24.87% of this function's abundance), Firmicutes (24.65%), and Proteobacteria (29.55%). Indeed, samples with high relative abundance of these phyla tended to have higher abundance of this subpathway. Similarly, examining the shares of pyruvate metabolism (a subpathway of carbohydrate metabolism) attributed to genera from the phylum Proteobacteria revealed a specifically large attribution of the genus *Neisseria*, a prominent

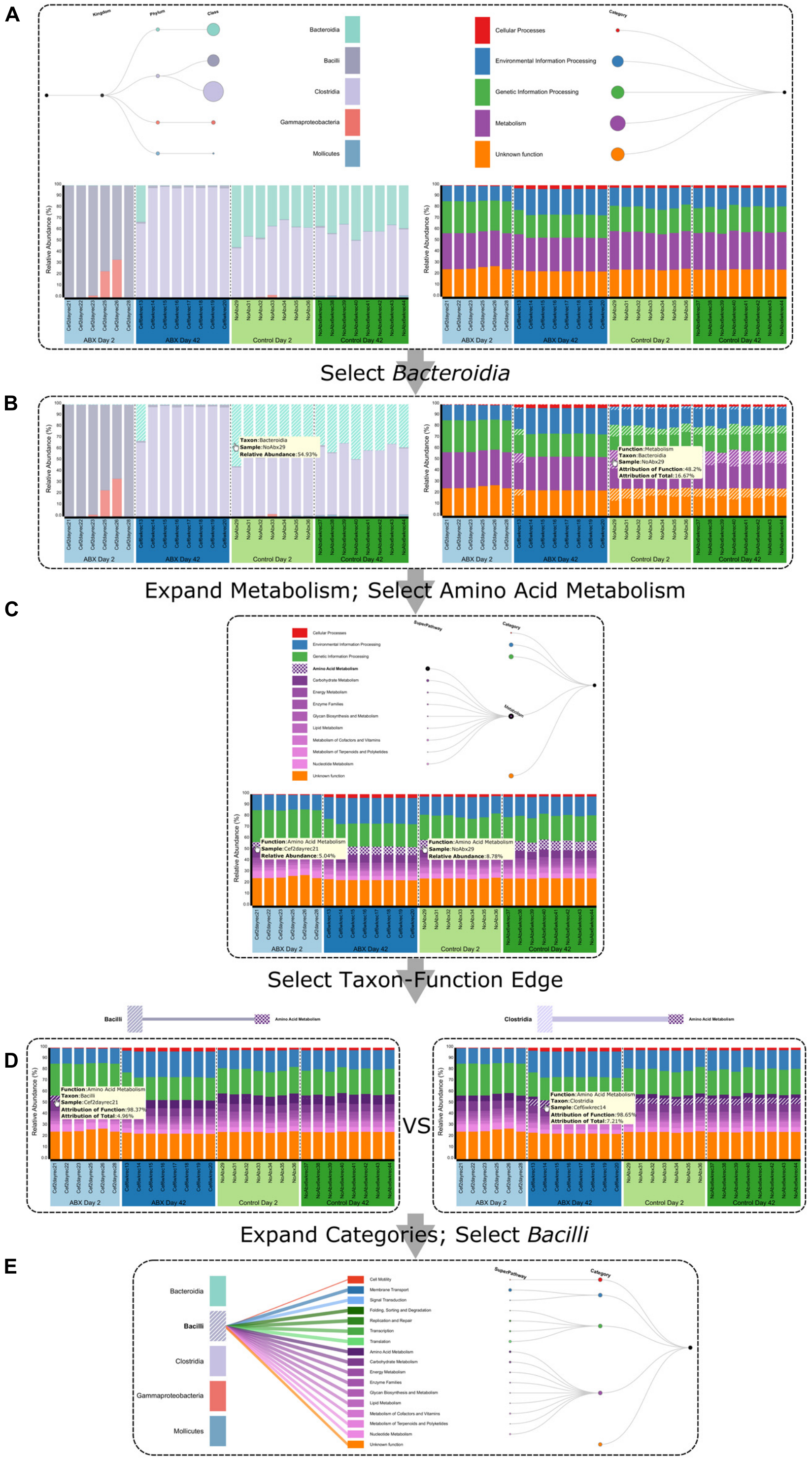


FIGURE 4 | Using BURRITO to visualize the functional impacts of microbiome shifts in response to antibiotics. **(A)** An overall view of BURRITO's display for this dataset of 29 mouse cecal samples. **(B)** Selecting a taxon (e.g., *Bacteroidia*) displays the share of each function attributed to this taxon. Tooltips provide exact attribution values. **(C)** Expanding functional resolution and clicking on a given function (e.g., Amino Acid Metabolism) highlights this function in each sample. Tooltips provide exact abundance values. **(D)** Once a function is selected, taxon-function edges in the control panel can be clicked, displaying the average share of the function attributed to the selected taxon and the sample-specific share of the function attributed to the selected taxon. **(E)** Edges connecting a given taxon to all encoded functions provide information about the average share of each function attributed to this taxon.

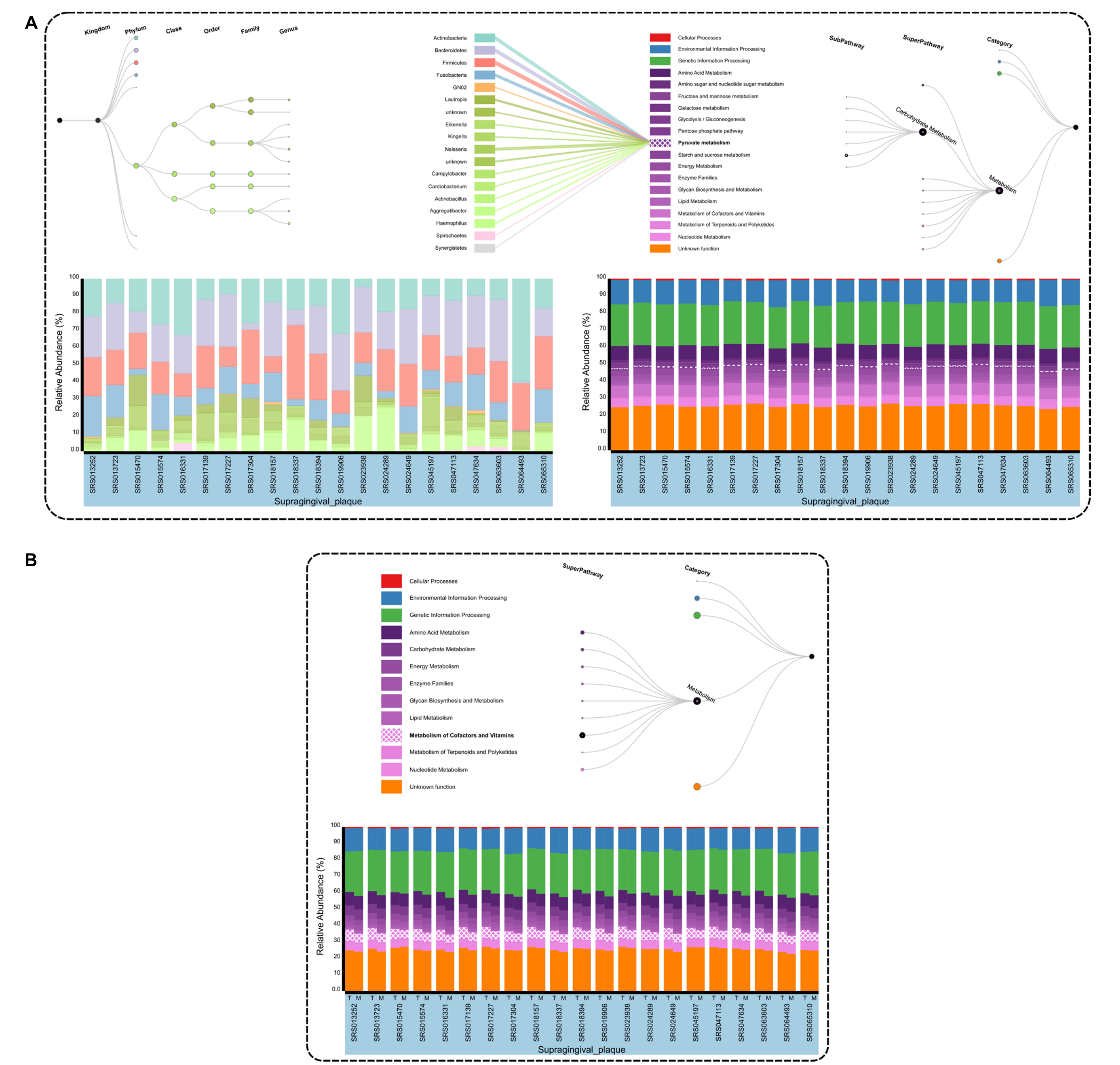


FIGURE 5 | Taxa–function relationships across supragingival plaque samples from the Human Microbiome Project. **(A)** Taxa and functions can be explored at different levels, using the trees in the control panel to expand taxa and functions of interest. This allows users to examine shares of specific functions that are attributed to specific taxa at high resolution. Edges in the control panel display information about the average share of a function attributable to each taxon across samples (with exact values provided via tooltips). **(B)** Inferred taxa-based functional abundances (linked to 16S rRNA taxonomic data, T) can be displayed alongside measured functional relative abundance data obtained from metagenomic shotgun sequencing (M) for easy comparison.

acid producer in the oral microbiome (**Figure 5A**). This is consistent with the capacity of strains from the genus *Neisseria* to metabolize lactate (via reactions included in the pyruvate metabolism pathway) (Hoshino and Araya, 1980; McLean et al., 2012).

Finally, **Figure 5B** demonstrates how BURRITO can also be used to compare such amplicon-based inferred functional profiles with functional abundance profiles obtained directly from shotgun metagenomic sequencing (when such functional profiles are provided as an additional input file). As expected, shotgun metagenomic-based profiles are generally in agreement with taxa-based inferred functional profiles, yet some differences can be observed, for example in the abundance of genes related to the metabolism of cofactors and vitamins (highlighted).

CONCLUSION

BURRITO is a web-based tool that addresses a major gap in currently available visualization tools for microbial ecology research. Studies in this field typically analyze, explore, and visualize taxonomic and functional abundances separately and fail to account for the interdependence between the two, potentially due to a shortage of tools that support simultaneous and integrative study of taxonomic and functional profiles. Our tool enables data exploration and hypothesis generation based on the attribution of functional abundances to specific taxa, providing a novel view into microbiome variation and dynamics.

AVAILABILITY AND REQUIREMENTS

1. **Project name:** BURRITO
2. **Project home page:** http://elbo.gs.washington.edu/software_burrito.html
3. **Operating system:** Platform independent
4. **Programming language:** HTML 5, JavaScript, D3
5. **Other requirements:** None
6. **Compatible browsers:** Chrome version 61.0+, Firefox version 53.0+, Opera version 48.0+, Safari 11.0+ (PNG exports not supported in Safari)
7. **License:** Licensed under the GNU General Public License v3.0 (available at https://github.com/borenstein-lab/Burrito).

AUTHOR CONTRIBUTIONS

CM, AE, CN, and EB conceived and designed this study, wrote the paper; CM, AE, CN, and WG-M implemented a preliminary version of BURRITO; CM, AE, and CN completed the implementation of BURRITO. All authors read and approved this manuscript.

FUNDING

CM was supported in part by "Interdisciplinary Training in Genomic Sciences" National Human Genome Research Institute T32 HG00035. CN and WG-M were supported in part by a National Science Foundation (NSF) IGERT DGE-1258485 fellowship. This work was supported in part by New Innovator Award DP2 AT007802-01 to EB.

ACKNOWLEDGMENTS

A preliminary version of BURRITO was implemented as an assignment in a Data Visualization course instructed by Jeffrey Heer and we thank him and the course TAs for their feedback. We also thank all members of the Borenstein lab for helpful discussions and software testing, and Casey Theriot for the data displayed in one of the case studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00365/full#supplementary-material

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Antunes, L. C. M., Han, J., Ferreira, R. B. R., Lolic, P., Borchers, C. H., and Finlay, B. B. (2011). Effect of antibiotic treatment on the intestinal metabolome. *Antimicrob. Agents Chemother.* 55, 1494–1503. doi: 10.1128/AAC.01664-10

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703. doi: 10.1016/j.chom.2015.04.004

Carr, R., Shen-Orr, S. S., and Borenstein, E. (2013). Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS Comput. Biol.* 9:e1003292. doi: 10.1371/journal.pcbi.1003292

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Greenblum, S., Carr, R., and Borenstein, E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160, 583–594. doi: 10.1016/j.cell.2014.12.038

Hoshino, E., and Araya, A. (1980). Lactate degradation by polysaccharide-producing Neisseria isolated from human dental plaque. *Arch. Oral Biol.* 25, 211–212. doi: 10.1016/0003-9969(80)90023-0

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

Jenior, M. L., Leslie, J. L., Young, V. B., and Schloss, P. D. (2017). *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. *mSystems* 2:e00063-17. doi: 10.1128/mSystems.00063-17

Jump, R. L. P., Polinkovsky, A., Hurless, K., Sitzlar, B., Eckart, K., Tomas, M., et al. (2014). Metabolomics analysis identifies intestinal microbiota-derived biomarkers of colonization resistance in clindamycin-treated mice. *PLoS One* 9:e101267. doi: 10.1371/journal.pone.0101267

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070

Kuntal, B. K., Ghosh, T. S., and Mande, S. S. (2013). Community-analyzer: a platform for visualizing and comparing microbial community structure across microbiomes. *Genomics* 102, 409–418. doi: 10.1016/j.ygeno.2013.08.004

Langille, M. G. I., 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

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature* 550, 61–66. doi: 10.1038/nature23889

Manor, O., and Borenstein, E. (2017a). Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. *Microbiome* 5:19. doi: 10.1186/s40168-017-0231-4

Manor, O., and Borenstein, E. (2017b). Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. *Cell Host Microbe* 21, 254–267. doi: 10.1016/j.chom.2016.12.014

McLean, J. S., Fansler, S. J., Majors, P. D., McAteer, K., Allen, L. Z., Shirtliff, M. E., et al. (2012). Identifying low pH active and lactate-utilizing taxa within oral microbiome communities from healthy children using stable isotope probing techniques. *PLoS One* 7:e32219. doi: 10.1371/journal.pone.0032219

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Barnabas, B., Blakesley, R., et al. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59–64. doi: 10.1038/nature13786

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A. H., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535, 376–381. doi: 10.1038/nature18646

Robertson, C. E., Harris, J. K., Wagner, B. D., Granger, D., Browne, K., Tatem, B., et al. (2013). Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics* 29, 3100–3101. doi: 10.1093/bioinformatics/btt526

Taxis, T. M., Wolff, S., Gregg, S. J., Minton, N. O., Zhang, C., Dai, J., et al. (2015). The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Res.* 43, 9600–9612. doi: 10.1093/nar/gkv973

Theriot, C. M., Koenigsknecht, M. J., Carlson, P. E. Jr., 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

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

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16. doi: 10.1186/2047-217X-2-16

Wang, Y., Xu, L., Gu, Y. Q., and Coleman-Derr, D. (2016). MetaCoMET: a web platform for discovery and visualization of the core microbiome. *Bioinformatics* 2:btw507. doi: 10.1093/bioinformatics/btw507

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569. doi: 10.1126/science.aad3369

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 Microbiology

TECHNOLOGY REPORT
published: 24 January 2018
doi: 10.3389/fmicb.2018.00036

'TIME': A Web Application for Obtaining Insights into Microbial Ecology Using Longitudinal Microbiome Data

Krishanu D. Baksi[1†], *Bhusan K. Kuntal*[1,2†] *and Sharmila S. Mande*[1*]

[1] *Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune, India,* [2] *Academy of Scientific and Innovative Research, CSIR-National Chemical Laboratory Campus, Pune, India*

Realization of the importance of microbiome studies, coupled with the decreasing sequencing cost, has led to the exponential growth of microbiome data. A number of these microbiome studies have focused on understanding changes in the microbial community over time. Such longitudinal microbiome studies have the potential to offer unique insights pertaining to the microbial social networks as well as their responses to perturbations. In this communication, we introduce a web based framework called 'TIME' (Temporal Insights into Microbial Ecology'), developed specifically to obtain meaningful insights from microbiome time series data. The TIME web-server is designed to accept a wide range of popular formats as input with options to preprocess and filter the data. Multiple samples, defined by a series of longitudinal time points along with their metadata information, can be compared in order to interactively visualize the temporal variations. In addition to standard microbiome data analytics, the web server implements popular time series analysis methods like Dynamic time warping, Granger causality and Dickey Fuller test to generate interactive layouts for facilitating easy biological inferences. Apart from this, a new metric for comparing metagenomic time series data has been introduced to effectively visualize the similarities/differences in the trends of the resident microbial groups. Augmenting the visualizations with the stationarity information pertaining to the microbial groups is utilized to predict the microbial competition as well as community structure. Additionally, the 'causality graph analysis' module incorporated in TIME allows predicting taxa that might have a higher influence on community structure in different conditions. TIME also allows users to easily identify potential taxonomic markers from a longitudinal microbiome analysis. We illustrate the utility of the web-server features on a few published time series microbiome data and demonstrate the ease with which it can be used to perform complex analysis.

Keywords: time series, microbiome, community state, visualization, clustering, Granger causality algorithm, web server

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Blake Hanson,
University of Texas Health Science Center at Houston, United States
Christine Nardini,
Partner Institute for Computational Biology, China
Karoline Faust,
KU Leuven, Belgium

****Correspondence:***
Sharmila S. Mande
sharmila.mande@tcs.com

†These authors have contributed equally to this work.

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

Received: *15 November 2017*
Accepted: *09 January 2018*
Published: *24 January 2018*

Citation:
Baksi KD, Kuntal BK and Mande SS (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

INTRODUCTION

Recent advances in high throughput next generation sequencing technologies and emergence of the field of metagenomics have helped in profiling not only the entire microbial groups in various environment(s), but also enabled cross sectional view of the sample(s) in a longitudinal time scale. While a cross sectional study design aims at comparisons of sample(s) at a single time point, longitudinal studies conduct several observations of the same sample(s) over a regular/irregular time intervals. A cross sectional study can provide insights regarding the differential abundances of the resident microbes across various states which is likely to be an indicator of potentially important biomarkers (Ghosh et al., 2014; Cameron et al., 2017). However, in order to obtain deeper understanding of the inter dependencies as well as periodic patterns and temporal variations in the microbial community, it is essential to perform a longitudinal study (Secrier and Schneider, 2014).

Temporal variation in microbial abundances play a critical role in influencing human health. For example, changes in microbial diversity are known to be associated with flu, seasonal allergies, as well as lifestyle disorders like diabetes and obesity (Hartstra et al., 2015; Riiser, 2015). The decreasing cost per mega-base of sequencing has enabled increased number of such large scale longitudinal metagenomic projects from diverse environments (Caporaso et al., 2011; Parsons et al., 2012; Kato et al., 2015). A number of studies have also concentrated in analyzing the changes in normal human microbiota after a perturbation event like administration of antibiotics (Dethlefsen and Relman, 2011). Unlike the cross sectional studies, the longitudinal microbiome studies have opened a new avenue for understanding the importance of causality analysis and networks based inferences on longitudinal time series microbiome data (Faust et al., 2015). New insights have also been obtained relating to differences in the stability of microbiomes across various environments (Shade et al., 2012). Another study has also elaborated the importance of stationarity analysis and its relation to microbial competition (David et al., 2014).

With the increase in number of microbiome projects, various tools and platforms have been developed for analysis of cross sectional microbiome data (Caporaso et al., 2010; Arndt et al., 2012; Kuntal et al., 2013; McMurdie and Holmes, 2013; Parks et al., 2014; Dhariwal et al., 2017; Kuntal and Mande, 2017). However, most of these tools cannot be utilized for understanding the temporal dynamics of microbial communities obtained from longitudinal studies. The available tools for time series microbiome data analysis are focused for a particular purpose or are implemented as library specific to a software platform (Bucci et al., 2016) which is difficult for biologists inexperienced in programming. While tools like Time-searcher (Hochheiser and Shneiderman, 2004) have options for visualizing any time series data, they have limited functionalities. STEM (Ernst and Bar-Joseph, 2006), TimeClust (Magni et al., 2008) and GATE (MacArthur et al., 2010), developed with a focus on microarray time series data, also cannot be used for time series microbiome data.

In order to obtain meaningful insights from microbiome time series data, we have developed a user friendly GUI web application, called 'TIME: (Temporal Insights into Microbial Ecology') publicly available at https://web.rniapps.net/time. 'TIME' allows users to upload data and perform analysis by selecting any desired workflow(s). Each workflow is carefully designed to address a biologically relevant question. These analyses include clustering similar taxa based on their temporal behavior, generating causality based inference networks, identification of time point similarities, etc. A new method for clustering time series data is also introduced and implemented in the platform. 'TIME' uses powerful visualization techniques coupled with interactive 'on the fly' analyses to assist obtaining meaningful inferences from microbiome time series data. Visual data mining and analysis of large time series datasets can be easily performed using this tool, thereby making it convenient for biologists to focus more on the results rather than implementation. 'TIME' intends to complement the existing metagenomic analysis tools and incorporate a suite of techniques that are suitable for microbiome time series analysis.

RESULTS

The 'TIME' Interface and the Workflows

A few time series microbiome studies have sampled data over a reasonably sized longitudinal span from individual(s) or environment(s) (Caporaso et al., 2011). Some of these time series datasets may consist of several short sampling stretches spanning over a long time period (Dethlefsen and Relman, 2011). The 'TIME' interface is designed to easily input user data in various formats (described in the "Materials and Methods" section) for visualization and analysis of time series microbiome data. Once the data is uploaded, a summary plot of the richness and diversity of microbial groups at each phylogenetic level is displayed. Following this, a user may proceed analyzing the data step by step selecting a workflow targeted for a specific time series analysis. Various workflows along with their biological implications are discussed below:

Workflow-1

Identify abundance based variations in taxonomic groups over time

The first and foremost step in any time series analysis pertains to visualization of temporal trends of the constituent entities (for example taxonomic groups in a microbial ecosystem). This workflow can be used to visualize and identify high, medium, and low abundant taxa. In addition, it allows identification of 'core,' 'persistent,' and 'transient' microbial groups which serve as important characteristic constituents of the ecosystem. The core microbiome refers to those taxa which are present across all time-points. On the other hand, the persistent microbiota refers to the ones that are present across extended time points, but not in all. In contrast, the transient group comprises of those sets of taxa which show frequent trends of appearance and disappearance. It should be noted that although the threshold parameters used for defining the 'core,' 'transient,' and 'persistent' have been taken from a previously reported study (Caporaso et al., 2011), they are prone to biases due to sequencing depth.

Workflow-2

Compare temporal trends between selected taxa

Analysis of time series data often requires trend comparison of a custom set of taxonomic groups. For example, the group may be a set of taxa previously known to show a characteristic behavior. The current workflow allows easy graphical comparison of two or more user selected taxa over the sampled timeline. TIME also allows comparison of trends in microbial abundance using a simple 'select and plot' operation, wherein users can choose microbial taxa (at a specified taxonomic level) using a simple auto-complete search or from a dropdown selection. One or more taxa can then be appended to or removed from the existing plot, thereby providing an easy way to study a selected set of microbes. Most of the microbiome datasets consist of sets of highly abundant as well as rare taxa. This poses a major problem while plotting and visualizing multiple taxa together in a single plot, with highly abundant taxa dominating the scale, thereby making it difficult to decipher patterns for the rare groups. In case the selected taxa have different abundance scales, users can utilize the 'log scaling' option for comparing their trends. Particular time stretches of interest can also be zoomed for in depth analysis.

Workflow-3

Identify temporally stable/unstable taxa

One of the important steps in time series analysis pertains to identification of stationary entities corresponding to the ones which have mean, standard deviation and variance constant over time. In microbial time series studies, identification of stationary taxa is especially crucial to detect inter microbial competition (David et al., 2014). As demonstrated in an earlier study (David et al., 2014), the presence of competition among the resident taxa is expected to cause sustained growth of some of them leading to their non-stationary behavior. Since in most cases the microbiota are in stable state, only a few taxonomic groups are expected to be non-stationary. A significant test of non-stationarity hence can be considered as a hint for a restoring force governing bacterial dynamics. However, fluctuations due to diet and environment may also affect stationarity of taxa and hence a cautious interpretation of results is required. Additionally, the similarities in phylogeny of non-stationary taxa may also provide clues pertaining to resource competition as genetically similar taxa are more likely to exhibit resource competition.

Workflow-4

Identify variations in taxonomic groups between two time ranges

Time series experiments involving perturbation events (like administration of antibiotics) are likely to disrupt the microbial community structure. In such analysis, it might be of interest to identify and visualize the exact temporal effect of the perturbation on the resident microbial groups. This workflow allows identification of taxa which undergo noticeable changes between two selected time ranges along with statistical inferences. Taxonomic behaviors like gradual increase or decrease in abundances can be easily inferred from the tabular summary generated using this workflow.

Workflow-5a

Cluster groups of taxa having similar behavior over time

An important goal while analyzing microbial time series data pertains to identification of groups of taxa which show similar trends over a time stretch. Similar temporal behavior by different bacterial taxa could arise due to reasons like symbiotic relationship between two or more bacteria. On the other hand, it is also important to know which bacterial taxa behave in temporally opposite ways, since such behavior might be an indicator of some underlying interaction or competition among them. Taxonomic groups depicting similar behavior in a selected timeframe are identified using Dynamic Time Warping (DTW) algorithm (described under "Materials and Methods" section). The output can be visually explored using interactive tree and trend plots. Each branch of the tree corresponds to a set of taxa having similar time series trends. Users can select a branch (a group of taxa having similar temporal patterns) or an individual terminal node (taxa) from the tree and visually explore the time series trends using the assistive plot.

Workflow-5b

Explore pair-wise relationship among taxonomic groups

Visualization of correlation and other similarity indices between the resident taxonomic groups often helps to gather meaningful insights. This workflow allows users to select Pearson correlation or modified DTW (referred to as TIME-DTW) index and use it to generate heatmaps. Such heatmaps are useful for visual pattern mining and the corresponding distance metric can be exported for further advanced network analysis.

Workflow-6

Explore inter taxa interactions using causality network

The existence of a strong correlation in the abundance of two or more taxonomic groups across a time scale may not always be ascertained to causation. A recent study has utilized 'module networks' to understand causality relationships among bacteria (Lu et al., 2017). A causation event can be ascertained between two taxonomic groups when the past values of one taxon are observed to have some information about the future values of the other. This analysis is performed in 'TIME' using a Granger causality algorithm (described in details under "Materials and Methods" section). The global community behavior over the whole sampled timeline is captured using interactive causality networks and trend plots. Each node in the network can be queried for its causality using interactive operations. While right clicking on a node (corresponding to a taxon) highlights the nodes (or taxa) that are affected ('Granger caused') by it, left clicking on the same highlights the nodes (or taxa) responsible for affecting ('Granger causing') its temporal changes.

Workflow-7

Cluster time points based on similar community patterns

Many microbial time series datasets are often observed to have a typical composition of constituent entities which gives rise to seasonality or periodicity of microbial communities. These similarities and differences in the proportion of the constituent taxonomic groups give rise to 'community states' in the microbiome. Such 'community states' could be useful for obtaining insights into the microbial dynamics (Gajer et al., 2012). The interactive hybrid trend plot and heatmap generated using this workflow is useful for visualizing the temporal changes in the community structure.

Case Studies on Publicly Available Time Series Microbiome Data

We demonstrate the applicability and utility of each workflow using three publicly available time series microbiome datasets. 'Caporaso-Dataset' corresponds to a longitudinal metagenomic time series data of gut microbiome samples from an American healthy male and female subject, collected at regular intervals spanning a long time period (Caporaso et al., 2011). A second time series metagenomic dataset ('Dethlefsen-Dataset') corresponds to a study evaluating the effects two doses of antibiotic treatments on the gut microbiome of three adult American females (Dethlefsen and Relman, 2011). The third dataset ('Gajer-Dataset') corresponds to a temporal sampling of vaginal microbiome of 32 reproductive age women over a period of 16-weeks (Gajer et al., 2012). All the above datasets are pre-loaded into the 'TIME' application for users' convenience.

Case Study 1: Analysis of Microbial Perturbation from Microbiome Time Series Data

In order to demonstrate the applicability of 'TIME' in analysis of perturbation, 'Dethlefsen-Dataset' (antibiotic treatment) was selected and various relevant workflows were used for analysis. The 'Dethlefsen-Dataset' had an associated metadata mapping for the time points corresponding to the different states (for all three individuals – D, E, and F), namely before antibiotic treatment ('PreCp'), during the two doses ('FirstCp' and 'SecondCp'), the week immediate post the two treatments ('FirstWPC' and 'SecondWPC'), gap between the doses ('Interim') and the time points post treatment ('PostCp'). A drastic drop in diversity and richness specifically at the points of perturbation ('FirstWPC' and 'SecondWPC'), could be visually inferred using the diversity plots generated using TIME (**Figure 1A**). The Figure also shows the slow but incomplete recovery in diversity post perturbation, which is in line with the reported findings (Dethlefsen and Relman, 2011). The core taxa identified using 'Workflow-1' (at 'genus' level) also indicates an inter-individual variation among the three subjects (**Figure 1B**), with only four genera to be consistently common across all (*Bacteroides*, *Coprococcus*, *Roseburia,* and *Dorea*). In order to identify the taxa which are most affected by the antibiotic treatment on 'Sample E' (as a representative example), the Workflow-4 was employed after selecting two time stretches, namely, 'before Cp1' (Period 1 ranging from time point 0–59) and 'after Cp1' (Period 2 ranging from time point 65–124). This analysis identified the affected genera sorted by the log fold change in the mean abundances between period 1 and period 2. *Haemophilus, Butyrivibrio, Eubacterium, Turicibacter,* and *Parabacteroides* were identified to be the top five affected genera upon antibiotic treatment based of log-fold abundance (**Figure 2**) but none of them were found to be statistically significant (when evaluated with Wilcoxon Rank-Sum Test using *P*-values corrected for multiple testing). Subsequently, to gather a deeper insight into the pattern of the affected genera during perturbation or genera similarly affected during perturbation, Workflow-5a was used (selecting sample 'E,' time point as 'FirstCp' and a rare taxa cutoff of 0.5) to generate the DTW tree (**Figure 3A**). Visual inference of the tree revealed three clear clusters (**Figure 3A**), each of which were used to generate their corresponding trend plots (**Figure 3B**). While Cluster 2 seemed to contain genera whose abundance is most strongly decreased by antibiotic treatment, Cluster 1 contained the moderately affected ones. On the other hand, Cluster 3 consisted of genera which increased post perturbation, possibly due to the reduced abundances of taxa belonging to Clusters 1 and 2.

Case Study 2: Insights into Microbial Inter-Dependencies Using Causality Networks

To analyze the effect of stationary genera and its relation to causality, the female subject (at genus level) from 'Caporaso-Dataset' (the 6 months spanning time series sampling) was used to generate the causality network using Workflow-6 (keeping a rare taxa cutoff of 0.5). The non-stationary genera information was overlaid on the network using one of the features in TIME which highlights the corresponding names (**Figure 4**). While a majority of the genera were seen to be stationary, a few exhibited non-stationary behaviors, an observation similar to an earlier study on a different gut microbiome dataset (David et al., 2014). Further, the majority of the non-stationary genera belonged to the phylum *Firmicutes*, strengthening the hypothesis of phylum level (genetically similar) resource competition (David et al., 2014). However, owing to the complexity of the community interactions in a gut microbiome, further experimental validations are required to support this hypothesis. In order to infer the effect of a non-stationary genus on others, we chose two non-stationary genera nodes (*Faecalibacterium* and *Clostridium*) from the causality network. While *Faecalibacterium* is a well documented commensal gut bacterium, a number of species belonging to *Clostridium* are known to have several pathogenic effects on human. Right clicking on these nodes enables one to highlight the edges connecting the genera affected ('Granger caused') by them and correspondingly displays the trend plot of all the associated taxonomic groups. A quick look into the edge connections showed that most of the genera affected by *Faecalibacterium* are non-stationary as compared to the ones affected by *Clostridium*. This observation suggests a differential influence of one taxon over others.

Case Study 3: Importance of Time Series Community Analysis

Microbial communities in different body sites have been reported to exhibit differences in their compositions. These compositions are also known to change over time. For example, studies on temporal variation of human gut microbiome have reported the presence of periodic as well as non-periodic diversity patterns (Caporaso et al., 2011). Such similar temporal patterns arise due to a comparable microbial community composition across these time points. Workflow-7 of 'TIME' is dedicated to identify such community clusters and visualize their variations across the timeline. 'Caporaso-Dataset' and 'Gajer-Dataset' (corresponding to gut and vaginal time series microbiome, respectively) were used as a part of this analysis pipeline (results are summarized in **Figures 5**, **6**, respectively). In order to consider the effect of only the 'non-rare taxa' (taxa which occur in at least 70% of

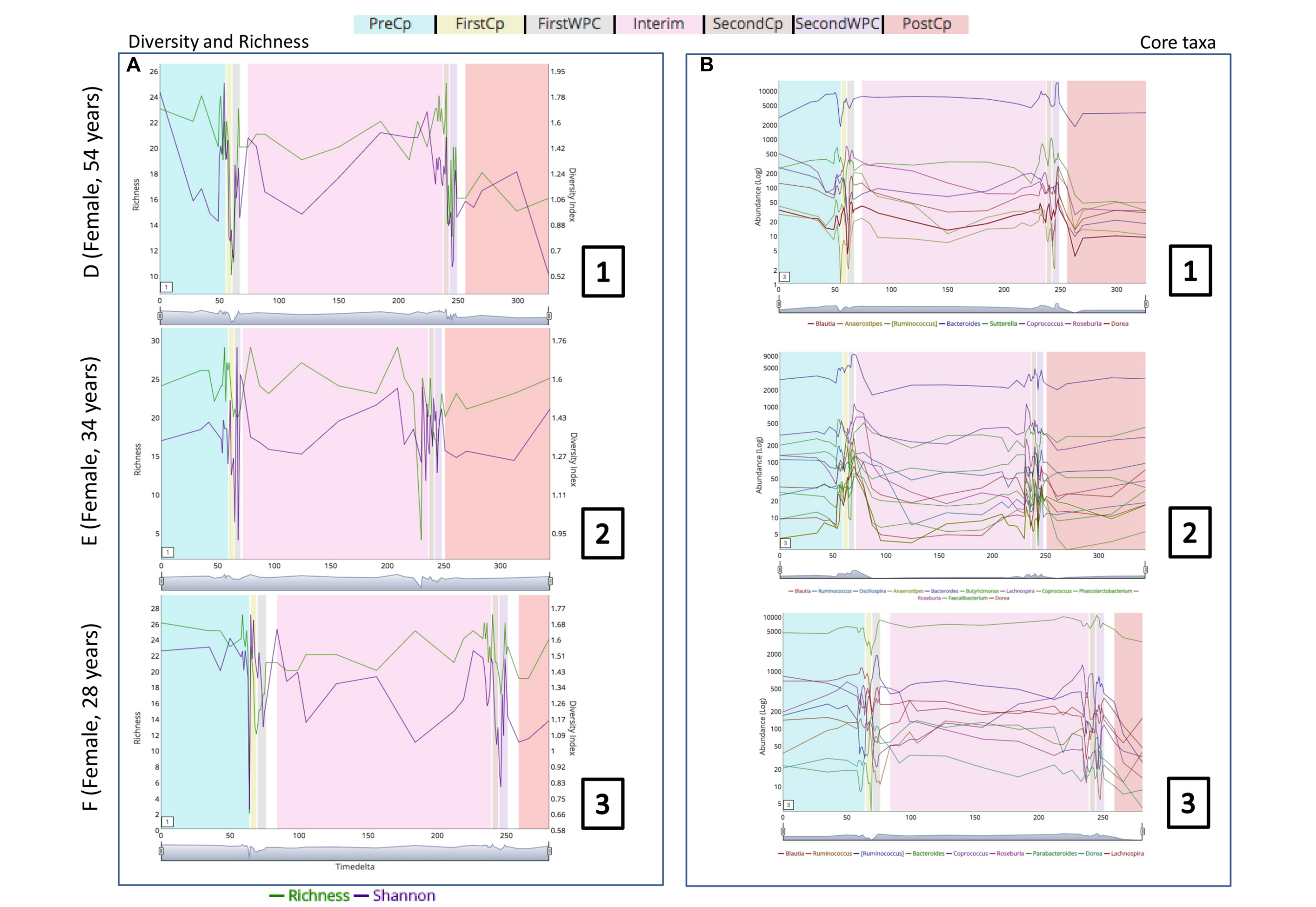
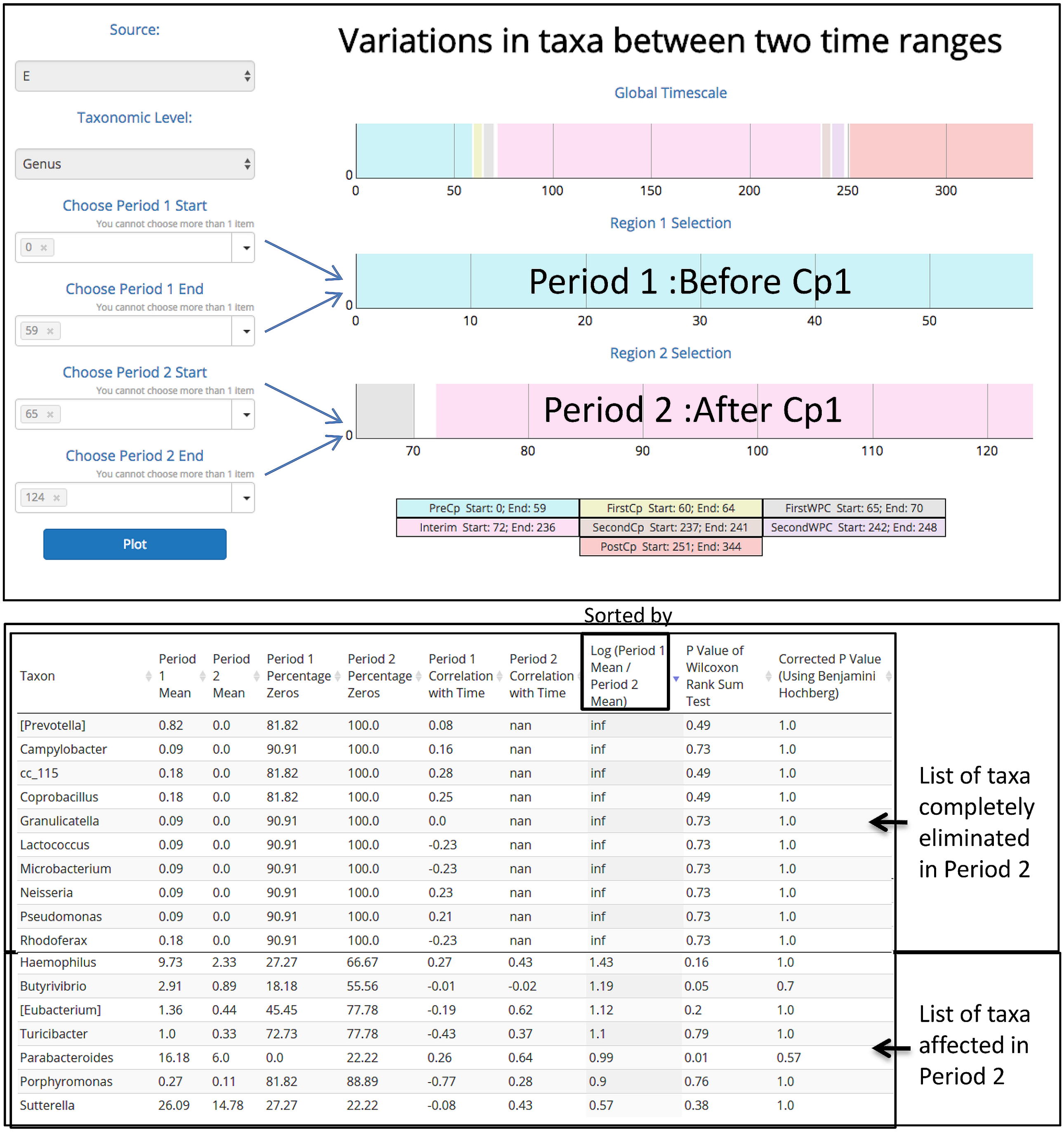


FIGURE 1 | (A) Changes in richness and diversity of the microbial genera across the three subjects (D, E, and F of 'Dethlefsen-Dataset' used in the case study) especially at the time points pertaining to antibiotic treatment ('FirstCp' and 'SecondCp'). **(B)** Trend plots of the core microbial genera in the three subjects show individual specific variations.

samples), a rare taxa cutoff of 0.7 was selected and a bi-directional clustering was done (for time points and taxa). Each of the two 'time-point clusters' (**Figure 5**) represent a group of time points having similar microbial distributions, called 'community states' (see section "Materials and Methods" for details). A comparison of the female and male gut microbiome time series ('Caporaso-Dataset') using the above workflow revealed a clear bias of one of the two 'community states' in the male (**Figure 5B**) while almost an equal distribution of the two 'community states' was found in the female (**Figure 5A**). In male, while the dominant cluster had mainly the genera *Bacteroides* and *Parabacteroides* as distinguishable marker, other genera namely *Prevotella* and *Campylobacter* were observed to be the prominent contributors of the less dominant cluster (**Figure 5B**). On the other hand, the female microbiome had one cluster prominently dominated by *Akkermansia,* with no single clearly dominant member in the other (**Figure 5A**). To explore community states in a different body site, vaginal microbiome from subject-1 of 'Gajer-Dataset' was considered and analyzed using Workflow-7. A clear periodic pattern in the 'community states' was observed (**Figure 6B**), probably due to the prominent changes in the menstrual cycle and related hormonal changes in reproductive age females. While one 'community state' showed a dominance of *Lactobacillus iners*, the other showed a dominance of *Atopobium*. The genera *Atopobium* is known to be associated with bacterial vaginosis, while lactic acid producing bacteria (like *L. iners*) are known to prevent pathogen colonization by creating an acidic environment (Gajer et al., 2012). The generated heatmap (**Figure 6A**) as well as trend comparison plot using Workflow-2 (**Figure 6C**) indicate an antagonistic behavior between the above two taxa.

DISCUSSION

The various workflows in 'TIME' allow visualizing time series microbiome data as well as analyzing them to obtain meaningful biological insights. It is to be noted that a few key points need to be considered before interpreting the generated outputs and building hypotheses based on such datasets (Weiss et al., 2017). For instance, the microbial abundance files used as input for analysis represent the count of clustered sequences (OTUs) across several time points corresponding to one or more



Source: E

Taxonomic Level: Genus

Choose Period 1 Start: 0

Choose Period 1 End: 59

Choose Period 2 Start: 65

Choose Period 2 End: 124

PreCp Start: 0; End: 59	FirstCp Start: 60; End: 64	FirstWPC Start: 65; End: 70
Interim Start: 72; End: 236	SecondCp Start: 237; End: 241	SecondWPC Start: 242; End: 248
	PostCp Start: 251; End: 344	

Sorted by

Taxon	Period 1 Mean	Period 2 Mean	Period 1 Percentage Zeros	Period 2 Percentage Zeros	Period 1 Correlation with Time	Period 2 Correlation with Time	Log (Period 1 Mean / Period 2 Mean)	P Value of Wilcoxon Rank Sum Test	Corrected P Value (Using Benjamini Hochberg)
[Prevotella]	0.82	0.0	81.82	100.0	0.08	nan	inf	0.49	1.0
Campylobacter	0.09	0.0	90.91	100.0	0.16	nan	inf	0.73	1.0
cc_115	0.18	0.0	81.82	100.0	0.28	nan	inf	0.49	1.0
Coprobacillus	0.18	0.0	81.82	100.0	0.25	nan	inf	0.49	1.0
Granulicatella	0.09	0.0	90.91	100.0	0.0	nan	inf	0.73	1.0
Lactococcus	0.09	0.0	90.91	100.0	-0.23	nan	inf	0.73	1.0
Microbacterium	0.09	0.0	90.91	100.0	-0.23	nan	inf	0.73	1.0
Neisseria	0.09	0.0	90.91	100.0	0.23	nan	inf	0.73	1.0
Pseudomonas	0.09	0.0	90.91	100.0	0.21	nan	inf	0.73	1.0
Rhodoferax	0.18	0.0	90.91	100.0	-0.23	nan	inf	0.73	1.0
Haemophilus	9.73	2.33	27.27	66.67	0.27	0.43	1.43	0.16	1.0
Butyrivibrio	2.91	0.89	18.18	55.56	-0.01	-0.02	1.19	0.05	0.7
[Eubacterium]	1.36	0.44	45.45	77.78	-0.19	0.62	1.12	0.2	1.0
Turicibacter	1.0	0.33	72.73	77.78	-0.43	0.37	1.1	0.79	1.0
Parabacteroides	16.18	6.0	0.0	22.22	0.26	0.64	0.99	0.01	0.57
Porphyromonas	0.27	0.11	81.82	88.89	-0.77	0.28	0.9	0.76	1.0
Sutterella	26.09	14.78	27.27	22.22	-0.08	0.43	0.57	0.38	1.0

List of taxa completely eliminated in Period 2

List of taxa affected in Period 2

FIGURE 2 | Demonstrating the utility of 'Workflow-4' in identifying the taxonomic groups completely eliminated and the ones mostly affected by antibiotic treatment during the first dosage period ('FirstCp' of 'Dethlefsen-Dataset' used in the case study). Two time periods ('BeforeCp' and 'AfterCp') were chosen based on the metadata.

sources (samples). Regardless of strict experimental designs, not all sources as well as time points are sampled/sequenced at similar depths due to sampling constraints as well as sequencing limitations. Hence, samples sequenced at lower depth may display biased diversity estimates and consequently affect the downstream analyses. For example, workflow 4 in TIME can predict differential abundant taxa between two time stretches with increased confidence if the sequencing depths are sufficiently high and even since samples with higher number of sequence will have better estimates of abundances. Similarly, if some time points are sampled deeper than the others, it makes interpretation of transient and rare taxa difficult (workflows 1

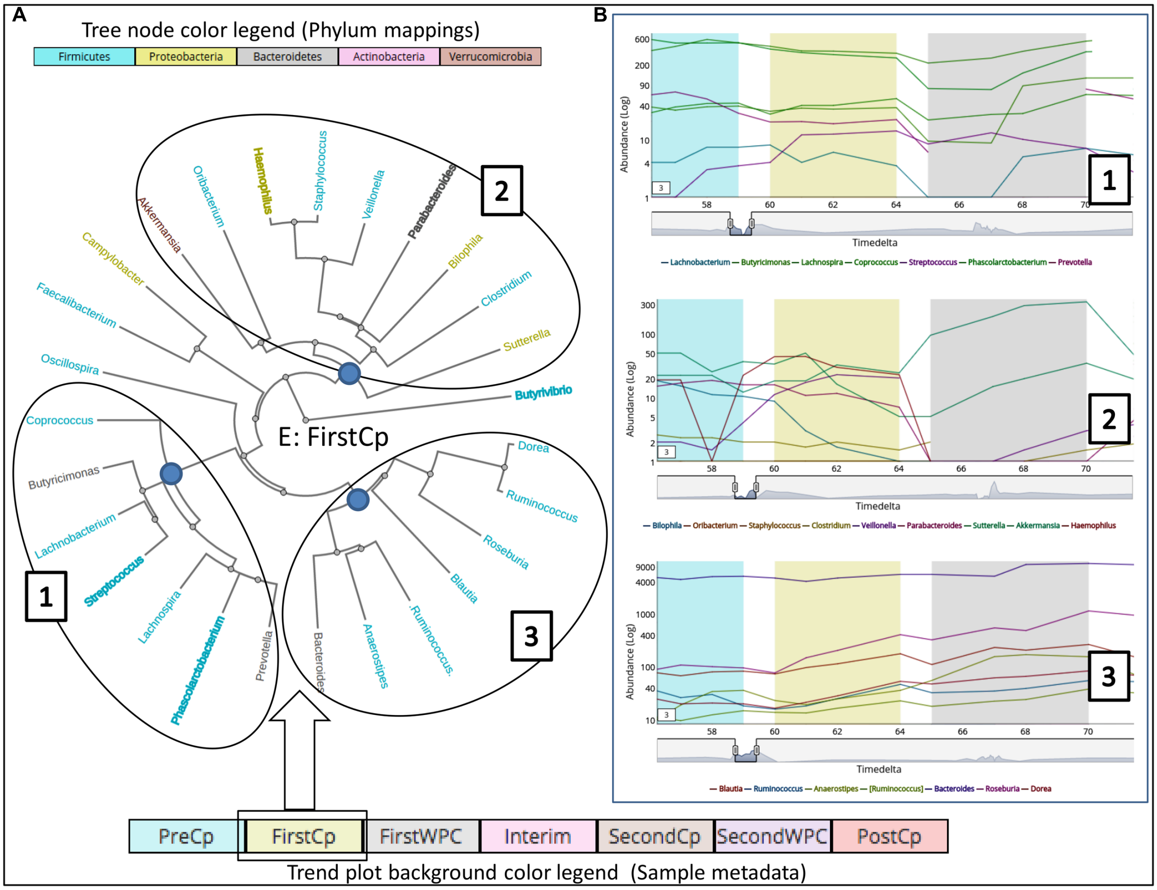


FIGURE 3 | Clustering of taxonomic groups based on their temporal trends during the antibiotic treatment ('FirstCp' of 'Dethlefsen-Dataset' used in the case study) on Subject 'E.' The **(A)** shows a tree (radial layout) with three clusters generated using DTW-distance metric in 'Workflow-5a.' The **(B)** shows the corresponding trend plots for the three clusters obtained by clicking on the root node of each cluster. The genera color labels (in the tree) correspond to their respective phyla as shown in the legend while bold labels indicate non-stationary taxa.

and 3) without normalization. In addition, presence of sparse OTUs represents uncertainty in counts owing to limitations in the sequencing detection ability (since they are below the detection threshold). A majority of microbiome studies consider either a relative normalization route (OTU counts scaled to proportions) or a rarefaction based normalization step (each sample is sub-sampled to an even depth), both of which are implemented in TIME for convenience. Use of rarefaction curves can provide guidance on choosing a suitable rarefaction depth for normalization and lower the false discovery rates (Weiss et al., 2017). However, it should be kept in mind that rarefying a data might impact a number of downstream analysis workflows due to removal of a subset of the data. Moreover, time series data involving perturbation events, if normalized using rarefaction, might subdue the effect of the perturbation itself. Relative normalization on the other hand, is also prone to create several artifacts (Stämmler et al., 2016). Both rarefied as well as relatively normalized data are compositional, therefore, fluctuations in abundance of one taxon might lead to spurious fluctuations in abundance of other taxa resulting in false correlations (Weiss et al., 2017). A lack of knowledge of absolute abundance can thus impact the interpretation of the results of the analyses. For example in workflow 3, although a taxon might change in abundance and appear to be non-stationary, it may actually be not changing but taxa around it may be changing in relative abundance. Moreover, relative abundance based approaches ignore the possibility that the altered abundance itself could be a key identifier of a disease state (Vandeputte et al., 2017). It may also be noted that both relative and absolute abundances are required for obtaining a comprehensive understanding of time series microbiome data (Props et al., 2017). Additionally, data obtained from appropriately designed experiments (e.g., using replicates for each time point) will increase confidence on the obtained results. Advanced experimental protocols have also been reported (Stämmler et al., 2016) which helps in normalizing the biases arising due to differential microbial loads across samples.

The incorporated Granger causality based interaction networks in 'TIME' provides a way to capture the overall global microbial community behavior and is ideal for datasets having evenly sampled time-points. Variations of Granger causality have been applied earlier to decipher ecological relationships

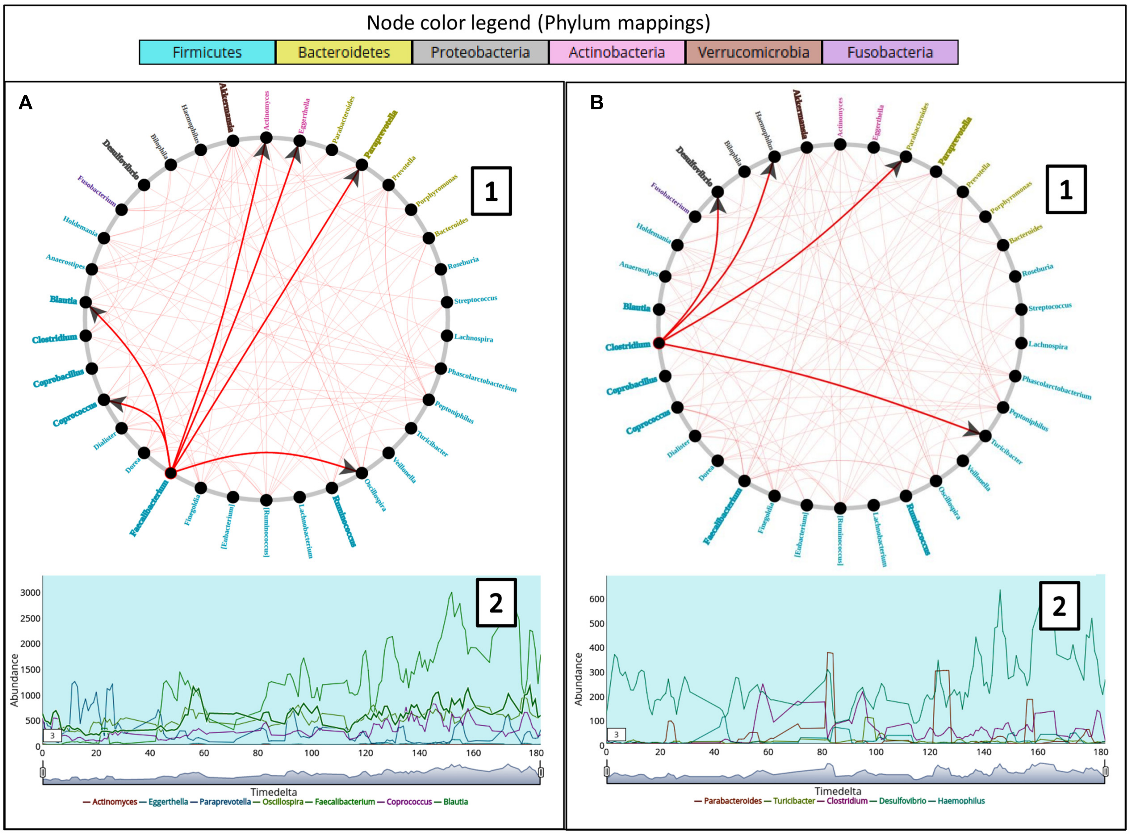


FIGURE 4 | (A,B) Represent two composite plots for Granger causality graphs (A1 and B1) and Trend plots (A2 and B2) corresponding to genera *Faecalibacterium* and *Clostridium* respectively. Granger causality graph (A1 and B1) for the constituent taxa in the female subject of 'Caporaso-Dataset' used in the case study generated using 'Workflow-6.' The trend plots (A2 and B2) for two genera namely *Faecalibacterium* and *Clostridium* along with the genera caused (or affected) by them are displayed below the corresponding circular graphs. The arrows in the graph represent the causality relationships between the source and target nodes. The genera color labels correspond to their respective phyla as shown in the legend while bold labels indicate non-stationary taxa.

(Detto et al., 2012) and in gene expression networks (Yang et al., 2017) with reasonable success. However, not all Granger causal interactions correctly predict biological causality and are merely statistical predictions. It should also be noted that such predictions do not provide explanations regarding the origin of the interactions and could be due to an indirect influence. For instance, one time series may be a strong predictor of another time series because both are shaped together by a common underlying cause. Hence, like any other statistical prediction, a cautious interpretation of each predicted interaction is required to be made before building any hypothesis. Incorporation of functional data like metabolic co-dependencies (Levy et al., 2015) might help to strengthen the basis of a predicted interaction.

CONCLUSION

The various workflows implemented in 'TIME' can help end users not only to perform a number of analyses, but also gain meaningful insights from the interactive visualizations. Analysis on a few well known publicly available datasets illustrate the utility of the options available in 'TIME.' For example, apart from obtaining information regarding the temporal effect of antibiotic treatment on human gut microbiome, 'TIME' could identify similarly perturbed groups of microbial genera. Additionally, the inter-microbial competition among the pathogens and commensals could also be inferred from the causality networks and stationarity analysis. In another example, the periodic changes in community structure of the vaginal microbiome were illustrated using the 'community state' analysis workflow. Although the scope of the case studies presented here is limited in this communication, the workflows can be further utilized to gain additional insights. We expect 'TIME' to be a valuable contribution in the field of microbial time series data analysis and visualization.

MATERIALS AND METHODS

'TIME' web application uses Python and JavaScript to execute the backend algorithms and for browser based data processing, respectively. We used the DyGraphs[1] (DyGraphs Java Script,

[1] http://dygraphs.com/

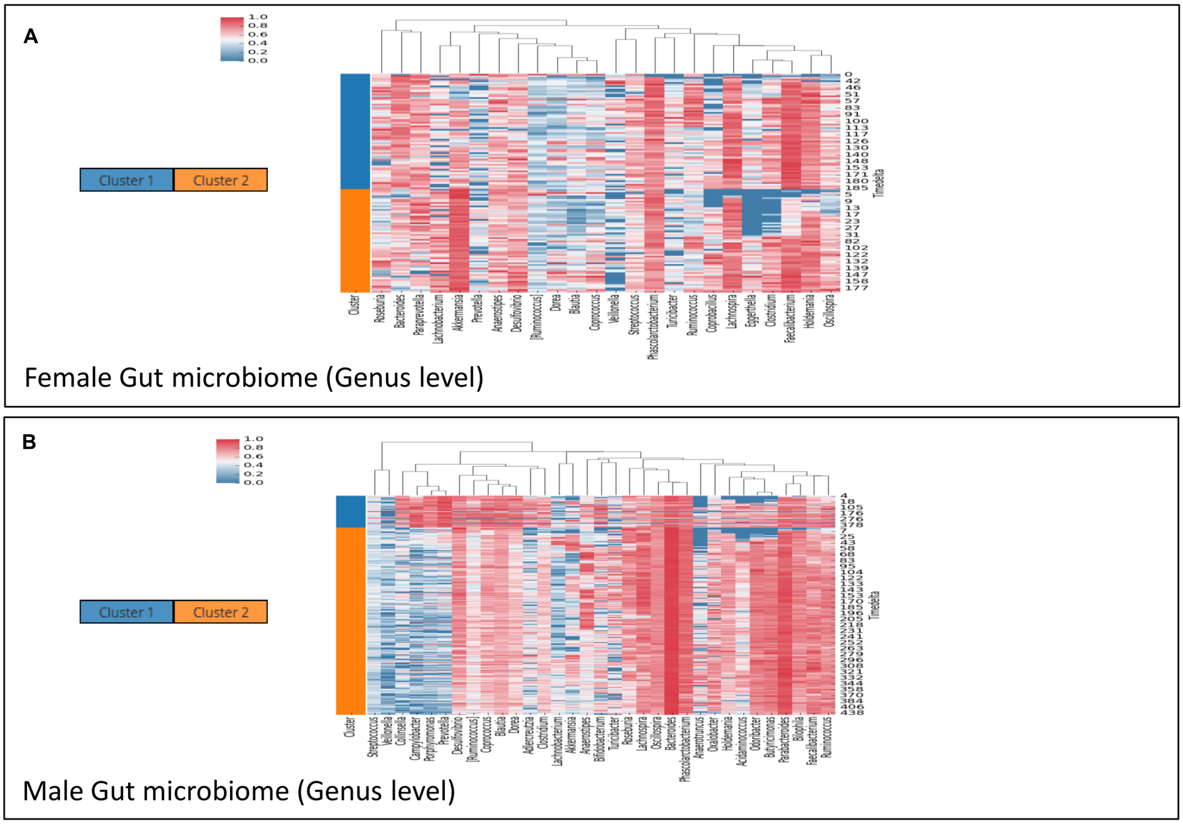


FIGURE 5 | Demonstrating the utility of 'Workflow-7' in gathering insights on community patterns in the female **(A)** and male **(B)** subject of 'Caporaso-Dataset' used in the case study. The heatmaps are clustered vertically based on taxa abundance and horizontally arranged according to the two 'community states' (represented as 'Cluster 1' and 'Cluster 2') identified by TIME.

2017) module for rendering time series line charts since it has the ability to handle large datasets seamlessly. Other visualizations are implemented using D3.js library (Bostock, 2017) with extensive interactive operations.

Input Format

User data (consisting of the microbial abundance table) along with the available metadata information can be incorporated in 'TIME' using a simple form. The abundance table can either be provided as a standard 'QIIME' output (Caporaso et al., 2010) or as a tab delimited file. The metadata file is required to have information related to the source of the microbiome sample, sample names (identical to the ones in the abundance file), time stamp information along with the sample condition for each time point. A detailed description of the input files is provided in the user manual (available in the website).

Normalization, Visual Exploration and Segregation of Microbiome Time Series Data

The microbial abundances obtained for analysis represents the count of clustered sequences belonging to the constituent taxa as operational taxonomic units (OTUs). The abundances of each OTU across different time points constitute the OTU abundance matrix. Restraints in sampling at multiple time points as well as sequencing errors result in unequal sequencing depths. 'TIME' provides methods to circumvent this limitation using either a proportion based or rarefaction based normalization. Rarefaction plots serve as one of the means to identify unequally sampled data points and subsequently can be used to normalize the OTU matrices such that all time points have similar counts. Users can generate a rarefaction curve for each metagenomic source by selecting either all the time points or a set of equidistant 5 or 10 time points. The generated curve can be used as a guide to select a suitable rarefaction normalization depth. Alternatively, users may proceed with absolute count data (without any normalization) or perform relative proportion based normalization. It is advisable to choose appropriate normalization method (refer to the "Discussion" section for more details).

The visual examination of the temporal trends is an important step in any time series analysis. In 'TIME,' all the taxa abundances at any particular taxonomic level can be viewed together as interactive line plots across the sampled timeline. An important challenge for carrying out such comparative microbial data analysis pertains to the problem of taxa abundances with

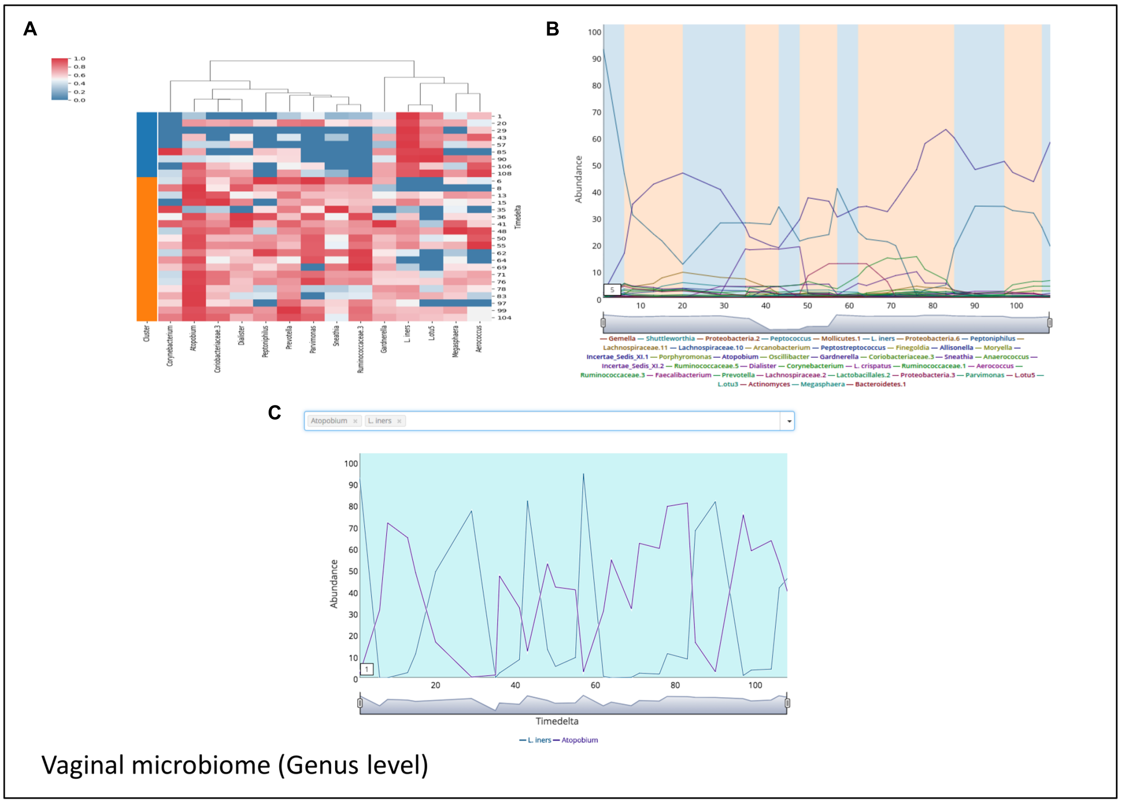


FIGURE 6 | Demonstrating the periodic microbial community patterns in the vaginal microbiome of 'Gajer-Dataset' used in the case study. **(A)** Heatmap clustered vertically based on taxa abundance and horizontally arranged based on the two community clusters (represented as 'Cluster 1' and 'Cluster 2') identified by TIME. **(B)** Trend plots of the constituent taxa with the plot background highlighted corresponding to the 'community state' affiliation (in 'blue' and 'orange') of the respective time points. **(C)** Demonstrating the antagonistic behavior between the two genera *Lactobacillus iners* and *Atopobium* in the vaginal microbiome dataset used in the case study generated using 'Workflow-2.'

different orders of magnitude (with some taxa having very high abundances and some having extremely low counts). In other words, it is difficult to visualize the trends of the lower abundant taxa owing to the dominant influence of the very high abundant ones on the plot. 'TIME' provides two ways to tackle this problem. While one uses 'quartile segregation,' the other utilizes 'log scaling.' In quartile segregation, the different taxa are grouped into four quartiles based on their abundance information which can be viewed separately. The very high abundant and the very low abundant taxonomic groups (or potential outliers) tend to occupy the top and bottom quartiles, respectively. The remaining quartile accommodates the taxonomic groups having the intermediate abundances. This makes sure that during visual exploration the temporal trends of the low abundance taxa do not get compressed (or dominated) by the trends of the very high abundant taxa. TIME also offers the option of log scaling the abundance values so that the trends of low abundant and that of high abundant taxa can be compared on the same plot. Additionally, the tool provides an option to view the core, persistent and the transient groups of bacteria which are reported to have distinct roles in microbial ecosystems (Caporaso et al., 2011). A taxon is considered to be persistent if it is observed in more than 20% of the time points, with at least 90% of these observations being consecutive. On the other hand, the transient taxa are those which are observed in at least 60% of the time points, with at most 75% of these observations being consecutive. However, TIME provides an option to modify these parameters (prevalence threshold and consecutive observations) for definition of core, persistent and transient in 'workflow 1' to accommodate differences in wide number of datasets. In addition to the above measures, the richness and diversity of the studied microbial communities are also calculated using well known indices. While richness of a microbiome denotes the unique number of constituent taxa present in each sample (at a time point), the diversity provides a measure of how evenly the taxonomic entities are distributed. Although diversity of a microbiome can be calculated using a number of ways, the widely accepted Shannon index for diversity (Shannon, 1948) has been implemented in 'TIME.'

$$Shannon\ \ Index\ \ = -\sum_{i=1}^{R} p_i\ \ \ln\ \ p_i$$

Where, p_{i} refers to the proportion of the abundance of the i^{th} taxon in the population consisting of 'R' taxa.

With implementation of each of the above methods and their corresponding visualizations, an interactive operation for selecting a 'subplot window' of the plotted timescale is presented. This feature enables users to graphically choose the start and end time points using simple mouse operations and visualize the selected time range. The 'subplot window' can then be dragged along the time-scale with a zoomed view. The moving average of a time series can also be specified using a text box available at the bottom corner of the plot window. This feature smoothes the short term fluctuations in the time series data and shows the overall trends (and cycles) across a longer timescale.

Identification of Stationary Taxonomic Groups

Stationarity of taxa in microbial time series data is important to understand inter-microbial competition (David et al., 2014). A taxon is considered stationary if its mean, variance, covariance, and autocorrelation are constant over time, due to the absence of a unit root process. A unit root process is said to be present in a time series if its autoregressive model has an estimated coefficient close to one. The presence of a unit root indicates that a perturbation in the value of the entity in the time series has a persistent impact on its future values and hence a cause of non-stationarity. The most commonly used method for calculating stationarity is the Augmented Dickey Fuller (ADF) Test. While the null hypothesis of the ADF Test is that there is a unit root process governing the dynamics of the entity (taxon), the alternate hypothesis states that there is no unit root. The ADF test statistic is a number, the more negative it is, the stronger will be the confidence with which the null hypothesis can be rejected. 'TIME' allows identification of microbial groups detected to be stationary and non-stationary and lists the same in a searchable table with options to export the results. The stationarity information corresponding to each taxon is also used to augment other plots in the tool along with the phylogeny information.

Generation of Inter-Microbial Causality Network

One of the important objectives in time series studies pertains to identifying causal relationships among entities. Causality aims to find the direct interactions between entities such that one entity can trigger/suppress or be triggered/suppressed by the other. It should be noted that causation should not be confused with correlation. For example, two taxa ('A' and 'B') in a microbiome time series dataset may be correlated, but may not have any causal relationships. Granger Causality (Granger, 1969) is one of the most well established statistical tests for checking causality among two time series. The basic premise of this method is that, if one variable causes another, the past values of the former must have some information about the future values of the latter (which is not available otherwise). For example in a microbiome time series data, if taxon A affects taxon B, the future values of taxon B can be better predicted using the past values of both taxa A and B, rather than using the past values of taxon B alone. In order to ascertain if taxon A influences taxon B, two regressions are performed. In one, past values of taxa A and B are used to predict the present values of taxon B. In the other, only the past values of taxon B is used to predict the present values of taxon B. If a significant increase of the goodness of fit of the former regression over the latter is observed, then taxon A is said to 'Granger cause' taxon B.

Since a typical microbiome time series data has more than two entities (taxa), 'Granger Lasso Causality' (Hlaváčková-Schindler and Pereverzyev, 2015) can be used to find causal relationships among all taxa. Thus, apart from the 'Pairwise Granger causality' (described above) for all possible taxa pairs, 'Granger-Lasso' method has also been implemented in 'TIME.' The LASSO (Least Absolute Shrinkage and selection operator) is one of the most well known and widely used methods for feature selection and regularization in machine learning. LASSO works by adding a regularizing penalty to the sum of squared errors. This objective function is minimized (by optimization) for estimating the values of the coefficients of regression (thus reducing the weightage/coefficients of the unimportant predictors), thereby finding the best set of predictors for every variable. Granger-Lasso utilizes the LASSO methodology for identifying causal relationships among all entities (taxa) in multivariate microbiome time series dataset (Arnold et al., 2007). Another option allowing selection of causality pairs predicted by both 'Pairwise Granger' and 'Granger-Lasso' is implemented for improved Granger Causality predictions. All these three methods are available in 'TIME' which can be finally used to generate a directed causality network.

Identifying Taxonomic Groups Having Similar Temporal Patterns

In time series datasets, it is not only important to evaluate temporal changes of different entities and the causal relationships among them, but also to identify entities which exhibit similar temporal patterns. The Euclidean distance based clustering of entities is unsuitable for identifying similar temporal patterns since this distance measure does not take into account the distortion across time series (Keogh and Ratanamahatana, 2005). In other words, the temporal behavior of two taxa which are out of phase is assigned a high value by Euclidian measure. On the other hand, Dynamic Time Warping (DTW) gives due importance to the phase displacement and obtains the optimal alignment between the two time series (Berndt and Clifford, 1994). DTW uses a dynamic programming based approach to align and score the similarity of the temporal patterns corresponding to two entities (taxa in the case of microbial time series). Since the DTW algorithm is relatively slow with a worst case time complexity of O(n^2), a modified DTW algorithm (Sakoe and Chiba, 1990) is implemented in 'TIME'. In this algorithm, a constraint is applied in such a way that a limited number of cells are evaluated during computing the cost matrix of the alignment, thereby making the overall computation process much faster (Salvador and Chan, 2007). 'TIME' uses the calculated pair-wise DTW distances among the different taxa for hierarchical clustering. The resulting dendrograms can be viewed as trees in standard or radial layouts. One of the limitations of the DTW distance pertains to the inability to interpret the distance score easily as it does not fall in a definite range. Therefore, it is desirable to have a modified score with a definitive range that can be universally interpreted. In order to achieve this, we introduce a new method for calculating the distance between two time series, called the 'TIME-DTW Distance,' In a microbial time series data, one taxon can have a difference of several orders of magnitude with another, but their time series may have similar overall shape. Thus, a standard normalization step is first applied to minimize such differences. Following this, the DTW distance is calculated and normalized by the average 'sum of the absolute difference' (SAD) between each time series and its 'mirror image' (Supplementary Material). The resultant value ('TIME-DTW distance') will hence always fall between a range of zero and one. 'TIME' allows an easy and interactive way to explore the results using a 'clustered heat map.' In addition to using the TIME DTW Distance as the measure of similarity/dissimilarity, the pairwise similarity between taxa can also be viewed using Pearson Correlation coefficient. The resulting heatmaps are hierarchically clustered based on their distances along the vertical axis, and taxonomic hierarchies along the horizontal axis.

Understanding Community Structure Based on Similarities across Time Points

Apart from understanding the temporal similarities among the resident entities (taxa), clustering of time points having similar entity distribution is expected to yield valuable insights regarding the microbial community dynamics. The identified time points having similar taxonomic distributions (i.e., phylotype proportions) can be considered as a 'community state' (Gajer et al., 2012). Jenson Shannon divergence (JSD) metric has been utilized earlier to identify such 'community state' in microbiome time series data (Gajer et al., 2012). In 'TIME,' a modification of the method is implemented to make it applicable for any microbiome time series. The taxa abundances are first normalized to generate probability distributions, which are then used to calculate the JSD among the different time points. Thus, a pairwise JSD matrix is obtained for all time points. Since, the *K*-medoids clustering algorithm (Jin and Han, 2011) is known to be robust to noise and outliers (as compared to *K*-means), it was utilized for clustering the time points using the generated JSD matrix. The number of clusters can be chosen by the user based on visual inspection. Along with clustering the different samples based on their microbial community structure, it is often useful and sometimes necessary to find the drivers of the cluster, i.e., the most dominant taxa among the clusters. Keeping this in view, 'TIME' also provides the option to view the (normalized) relative abundances of the taxa across different clusters and different time points as a heatmap, which helps in visual exploration and determination of the distinctive/driver taxa or groups of taxa.

AUTHOR CONTRIBUTIONS

KB, BK, and SM conceived the idea and designed the overall methodology. KB and BK implemented the algorithms and developed the web server. BK performed the case studies. BK, KB, and SM analyzed the results and drafted the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We acknowledge Dr. Swadha Anand and Dr. Anirban Dutta for their help during the course of this work. We also acknowledge Mr. Pramod Nanadikar and Mr. Govind Gopal for their help in deploying the application. BK is also an industry sponsored Ph.D. student at Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune 411008 (India) and would like to acknowledge the Academy of Scientific and Innovative Research (AcSIR) for its support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00036/full#supplementary-material

REFERENCES

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., et al. (2012). METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* 40, W88–W95. doi: 10.1093/nar/gks497

Arnold, A., Liu, Y., and Abe, N. (2007). "Temporal causal modeling with graphical granger methods," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '07*, (New York, NY: ACM), 66–75. doi: 10.1145/1281192.1281203

Berndt, D. J., and Clifford, J. (1994). "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining AAAIWS'94*, (Seattle, WA: AAAI Press), 359–370.

Bostock, M. (2017). *D3. JS - Data-Driven Documents*. Available at: https://d3js.org/ [accessed November 2, 2017].

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

Cameron, S. J. S., Lewis, K. E., Huws, S. A., Hegarty, M. J., Lewis, P. D., Pachebat, J. A., et al. (2017). A pilot study using metagenomic sequencing of the sputum microbiome suggests potential bacterial biomarkers for lung cancer. *PLOS ONE* 12:e0177062. doi: 10.1371/journal.pone.0177062

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

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

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

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. U.S.A.* 108(Suppl. 1), 4554–4561. doi: 10.1073/pnas.1000087107

Detto, M., Molini, A., Katul, G., Stoy, P., Palmroth, S., and Baldocchi, D. (2012). Causality and persistence in ecological systems: a nonparametric spectral granger causality approach. *Am. Nat.* 179, 524–535. doi: 10.1086/664628

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* doi: 10.1093/nar/gkx295 [Epub ahead of print].

DyGraphs Java Script (2017). Available at: http://dygraphs.com/ [accessed November 2, 2017].

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

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

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4, 132ra52. doi: 10.1126/scitranslmed.3003605

Ghosh, T. S., Gupta, S. S., Bhattacharya, T., Yadav, D., Barik, A., Chowdhury, A., et al. (2014). Gut microbiomes of Indian children of varying nutritional status. *PLOS ONE* 9:e95547. doi: 10.1371/journal.pone.0095547

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438. doi: 10.2307/1912791

Hartstra, A. V., Bouter, K. E. C., Bäckhed, F., and Nieuwdorp, M. (2015). Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* 38, 159–165. doi: 10.2337/dc14-0769

Hlaváčková-Schindler, K., and Pereverzyev, S. (2015). "Lasso granger causal models: some strategies and their efficiency for gene expression regulatory networks," in *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability Studies in Computational Intelligence*, eds T. Guy, M. Kárný, and D. Wolpert (Cham: Springer), 91–117. doi: 10.1007/978-3-319-15144-1_4

Hochheiser, H., and Shneiderman, B. (2004). Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Inf. Vis.* 3, 1–18. doi: 10.1057/palgrave.ivs.9500061

Jin, X., and Han, J. (2011). "K-medoids clustering," in *Encyclopedia of Machine Learning*, eds C. Sammut and G. I. Webb (Berlin: Springer), 564–565. doi: 10.1007/978-0-387-30164-8_426

Kato, H., Mori, H., Maruyama, F., Toyoda, A., Oshima, K., Endo, R., et al. (2015). Time-series metagenomic analysis reveals robustness of soil microbiome against chemical disturbance. *DNA Res.* 22, 413–424. doi: 10.1093/dnares/dsv023

Keogh, E., and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* 7, 358–386. doi: 10.1007/s10115-004-0154-9

Kuntal, B. K., Ghosh, T. S., and Mande, S. S. (2013). Community-analyzer: a platform for visualizing and comparing microbial community structure across microbiomes. *Genomics* 102, 409–418. doi: 10.1016/j.ygeno.2013.08.004

Kuntal, B. K., and Mande, S. S. (2017). Web-igloo: a web based platform for multivariate data visualization. *Bioinformatics* 33, 615–617. doi: 10.1093/bioinformatics/btw669

Levy, R., Carr, R., Kreimer, A., Freilich, S., and Borenstein, E. (2015). NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics* 16:164. doi: 10.1186/s12859-015-0588-y

Lu, Y., Zhou, X., and Nardini, C. (2017). Dissection of the module network implementation "LemonTree": enhancements towards applications in metagenomics and translation in autoimmune maladies. *Mol. Biosyst.* 13, 2083–2091. doi: 10.1039/c7mb00248c

MacArthur, B. D., Lachmann, A., Lemischka, I. R., and Ma'ayan, A. (2010). GATE: software for the analysis and visualization of high-dimensional time series expression data. *Bioinformatics* 26, 143–144. doi: 10.1093/bioinformatics/btp628

Magni, P., Ferrazzi, F., Sacchi, L., and Bellazzi, R. (2008). TimeClust: a clustering tool for gene expression time series. *Bioinformatics* 24, 430–432. doi: 10.1093/bioinformatics/btm605

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE* 8:e61217. doi: 10.1371/journal.pone.0061217

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Parsons, R. J., Breitbart, M., Lomas, M. W., and Carlson, C. A. (2012). Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J.* 6, 273–284. doi: 10.1038/ismej.2011.101

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

Riiser, A. (2015). The human microbiome, asthma, and allergy. *Allergy Asthma Clin. Immunol.* 11:35. doi: 10.1186/s13223-015-0102-0

Sakoe, H., and Chiba, S. (1990). "Dynamic programming algorithm optimization for spoken word recognition," in *Readings in Speech Recognition*, eds A. Waibel and K.-F. Lee (San Francisco, CA: Morgan Kaufmann Publishers), 159–165.

Salvador, S., and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* 11, 561–580.

Secrier, M., and Schneider, R. (2014). Visualizing time-related data in biology, a review. *Brief. Bioinform.* 15, 771–782. doi: 10.1093/bib/bbt021

Shade, A., Read, J. S., Youngblut, N. D., Fierer, N., Knight, R., Kratz, T. K., et al. (2012). Lake microbial communities are resilient after a whole-ecosystem disturbance. *ISME J.* 6, 2153–2167. doi: 10.1038/ismej.2012.56

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

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

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

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

Yang, G., Wang, L., and Wang, X. (2017). Reconstruction of complex directional networks with group lasso nonlinear conditional granger causality. *Sci. Rep.* 7:2991. doi: 10.1038/s41598-017-02762-5

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.

METHODS
published: 26 March 2018
doi: 10.3389/fmicb.2018.00522

A Bayesian Semiparametric Regression Model for Joint Analysis of Microbiome Data

Juhee Lee[1]* *and Marilou Sison-Mangus*[2]

[1] *Department of Applied Mathematics and Statistics, University of California, Santa Cruz, Santa Cruz, CA, United States,* [2] *Department of Ocean Sciences, University of California, Santa Cruz, Santa Cruz, CA, United States*

The successional dynamics of microbial communities are influenced by the synergistic interactions of physical and biological factors. In our motivating data, ocean microbiome samples were collected from the Santa Cruz Municipal Wharf, Monterey Bay at multiple time points and then 16S ribosomal RNA (rRNA) sequenced. We develop a Bayesian semiparametric regression model to investigate how microbial abundance and succession change with covarying physical and biological factors including algal bloom and domoic acid concentration level using 16S rRNA sequencing data. A generalized linear regression model is built using the Laplace prior, a sparse inducing prior, to improve estimation of covariate effects on mean abundances of microbial species represented by operational taxonomic units (OTUs). A nonparametric prior model is used to facilitate borrowing strength across OTUs, across samples and across time points. It flexibly estimates baseline mean abundances of OTUs and provides the basis for improved quantification of covariate effects. The proposed method does not require prior normalization of OTU counts to adjust differences in sample total counts. Instead, the normalization and estimation of covariate effects on OTU abundance are simultaneously carried out for joint analysis of all OTUs. Using simulation studies and a real data analysis, we demonstrate improved inference compared to an existing method.

Keywords: count data, Laplace prior, metagenomics, microbiome, regularizing prior, process convolution, negative binomial model, 16S ribosomal RNA sequencing

OPEN ACCESS

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Yanxun Xu,
Johns Hopkins University,
United States
Michael Pester,
Deutsche Sammlung von
Mikroorganismen und Zellkulturen
(DSMZ), Germany

Correspondence:
Juhee Lee
juheelee@soe.ucsc.edu

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

Received: *15 November 2018*
Accepted: *08 March 2018*
Published: *26 March 2018*

Citation:
Lee J and Sison-Mangus M (2018) A
Bayesian Semiparametric Regression
Model for Joint Analysis of
Microbiome Data.
Front. Microbiol. 9:522.
doi: 10.3389/fmicb.2018.00522

1. INTRODUCTION

Microbial communities are influenced by several factors whether they live in the host's guts or other occupied niches. Their successional dynamics could further change in response to perturbations of the host or of the surrounding environments (Turnbaugh et al., 2009; Needham and Fuhrman, 2016). Understanding how abiotic and biotic factors influence the dynamics of microbial communities is of great interest in the field of microbiome studies. Recent revolutionary advances in next-generation sequencing (NGS) technologies along with rapidly decreasing costs, have facilitated the accumulation of large datasets of 16S ribosomal RNA (rRNA) amplicon sequences across various disciplines such as medicine, biology, ecology, and environmental sciences (Woo et al., 2008). Sequencing data is usually pre-treated for quality filtering, noise removal and chimera checking through bioinformatics algorithms and the filtered sequences are clustered into Operational Taxonomic Units (OTUs), which represent similar organisms (microbial species) based on sequence homology (called OTU picking). An OTU abundance table is generated, recording counts for OTUs in samples. Further statistical data analyses are then performed using the OTU table to answer biological and ecological questions.

Analysis of huge NGS data is computationally expensive and challenging. One of the key challenges is the normalization of counts across samples. Total counts (often called library size or sequencing depth) may vastly vary across different samples due to technical reasons. Thus, observed counts are not directly comparable across samples and cannot be used as a measure of the abundance of an OTU. Normalized counts through rarefaction or relative frequencies are commonly used for easy comparison of OTU abundance across samples. However, such *ad hoc* normalization procedures have been criticized from a statistical perspective since using pre-normalized quantities may undermine the performance of downstream analysis (McMurdie and Holmes, 2014). Another challenge is high dimensionality and sparsity in OTU count data. A dataset typically includes hundreds or thousands of OTUs and a majority of them has zero or very low frequencies in most of samples. For example, **Figure 1A** illustrates a heatmap of OTU counts in our motivating dataset described in section 2.3. It shows that a majority of OTUs has very low counts (gray) in a sample, and the set of OTUs having large counts (blue) vary across samples. Due to such sparsity in data, borrowing strength across OTUs through joint analysis of all OTUs is crucial for improved inference. Recently, various statistical methods including Romero et al. (2014), Chen and Li (2016), Gibbons et al. (2017), and Zhang et al. (2017) have been developed for microbiome studies using NGS data. For example, Zhang et al. (2017) used a negative binomial mixed regression model to study interactions between the microbiome and host environmental/clinical factors. Random effects are used to induce correlation among samples from a group. Common to most of recent methods including Zhang et al. (2017) is separately analyzing each OTU at a time.

We develop a Bayesian semiparametric generalized linear regression model to study the effects of physical and biological factors on abundance of microbes. The proposed method performs mode-based normalization through a hierarchical model, which enables direct modeling of OTU counts. Furthermore, the hierarchical model facilitates borrowing strength between OTUs, between samples, and between time points through joint analysis and improves inference on the effects of covariates $\boldsymbol{X}$ on OTU abundance which are the parameters of our primary interest. Specifically, a negative binomial (NB) distribution parameterized by a mean parameter μ and an overdispersion parameter s is assumed for OTU counts. The NB distribution flexibly accommodates overdispersion often seen in NGS data and is commonly used as a robust alternative to a Poisson distribution (Anders and Huber, 2010). The expected count μ of an OTU is decomposed as a product of factors, a baseline mean count g and a nonnegative function $\eta(\boldsymbol{X})$ of covariates that describes their effects on the mean count. We use the log link function for $\eta(\boldsymbol{X})$ and assume that change in a covariate has a multiplicative effect on mean count, where the associated coefficient quantifies the size and direction of the effect. We consider a Laplace prior for the coefficients, a shrinkage prior that is essential in a high dimensional regression setting. Shrinkage priors in regression yield sparse point estimates of the coefficients, where many of the coefficients have values close to zero and few have large values. The sparse estimates improve out-of-sample prediction and produce more interpretable models (Park and Casella, 2008). In addition, shrinkage priors such as a Laplace prior in a regression problem mitigate potential problems by multicollinearity and yield improved coefficient estimates when covariates are high-dimensional and potentially highly correlated (Polson and Scott, 2012). For baseline mean counts, we develop a nonparametric model to combine all OTUs for joint analysis. Baseline mean counts may vary across samples and OTUs. Also, as in our motivating data for which samples were taken over time, there may be temporal dependence in baseline mean counts. To tackle the problem, we further decompose the baseline count g into sample size factor (r), OTU size factor (α_0), and OTU and time factor (α_t), that is, $g = r \times \alpha_0 \times \alpha_t$. Due to the overparametrization of the baseline mean abundance, individual factors are not identifiable. To avoid identifiability issues, we place the regularizing priors with mean constraints (Li et al., 2017) for sample size factor r and OTU size factor α_0. In addition, we model a temporal dependence structure between the baseline expected counts for an OTU through a convolutional Gaussian process (Higdon, 1998). The process convolution approach is often used as an alternative approach of the Gaussian process to construct a dependent process due to its efficient computation (Lee et al., 2005; Liang and Lee, 2014). Through simulation studies, we show that estimates of individual parameters r, α_0, and α_t are not fully interpretable under the proposed model, but baseline mean counts g are identifiable. The model also provides a posterior distribution of g for uncertainty quantification.

The rest of the paper is organized as follows. In section 2 we describe the proposed model and discuss the prior formulations and the resulting posterior inference. We perform simulation studies to assess the proposed model and perform comparison with an existing method that analyzes one OTU at a time. We then apply the proposed model to an ocean microbiome dataset. Section 3 presents the performance of the proposed model from the simulation experiment and the ocean microbime data. Section 4 concludes the paper with a discussion on limitations and possible extensions.

2. MATERIALS AND METHODS

2.1. Bayesian Semiparametric Regression Model

Suppose that samples are taken at n different time points, $0 \leq t_i \leq T$, $i = 1, \ldots, n$, and with K_i replicates at time point t_i. We consider count $y_{t_i,k,j}$ of OTU j in replicate k taken at time t_i, where $i = 1, \ldots, n$, $k = 1, \ldots, K_i$, and $j = 1, \ldots, J$. A sample is thus indexed by t_i and k. We let the total number of samples $N = \sum_{i=1}^{n} K_i$. Let $\boldsymbol{Y} = [y_{t_i,k,j}]$ denote the $N \times J$ matrix of counts, where $y_{t_i,k,j}$ is integer-valued and nonnegative. Also, suppose that covariates $\boldsymbol{X}_{t_i} = (X_{t_i,1}, \ldots, X_{t_i,P})'$ are recorded at

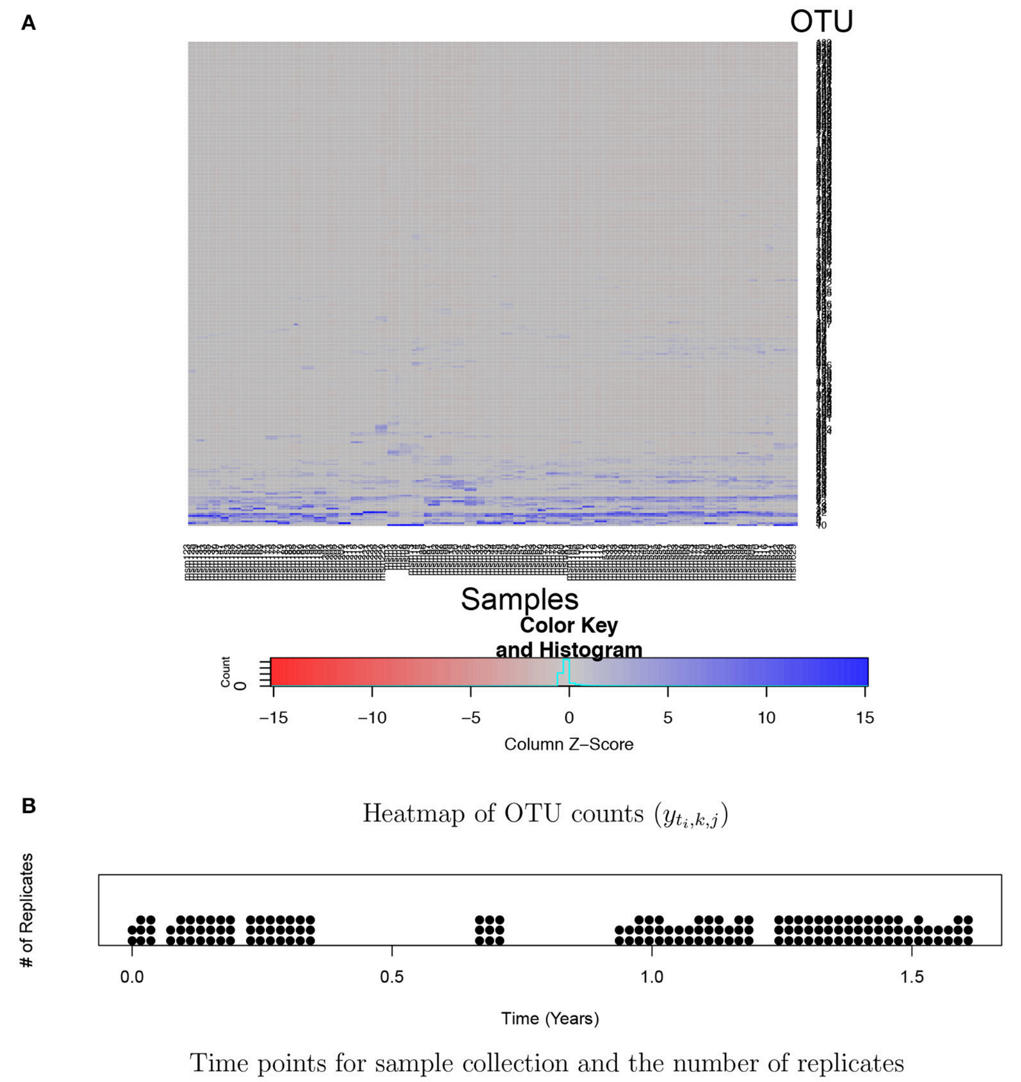


FIGURE 1 | Ocean microbiome data. **(A)** Heatmap of OTU counts ($y_{t_i,k,j}$). OTU and samples are in rows and columns, respectively. OTU counts are rescaled within a sample for better illustration. **(B)** 55 time points where ocean microbiome samples were collected are marked on the X-axis and the number of dots at a time point represents the number of replicates (K_i) at the time point.

time t_i. For example, covariates are physical and biological factors in our motivating data.

2.1.1. Sampling Model

Count data by NGS methods is often modeled through a Poisson distribution. The assumption under the Poisson distribution that the variance is equal to the mean is often too restrictive to accommodate overdispersion that variation in data exceeds the mean. The negative binomial (NB) distribution is a popular and convenient alternative to address the overdispersion problem and is widely recognized as a model that provides improved inference to NGS count data (for example, see Robinson and Smyth, 2007; Anders and Huber, 2010). A NB distribution can be characterized by mean and overdispersion parameters. We suppress index i for simpler notation and assume a NB model for count $y_{t,k,j}$ of OTU j in replicate k at time t,

$$y_{t,k,j} \overset{indep}{\sim} \text{NB}(\mu_{t,k,j},\ s_j), \tag{1}$$

where mean count $\mu_{t,k,j} > 0$ and overdispersion parameter $s_j > 0$. The model in Equation (1) implies that count of OTU j in replicate k at time t has mean $\text{E}(y_{t,k,j} \mid \mu_{t,k,j}) = \mu_{t,k,j}$ and variance $\text{Var}(y_{t,k,j} \mid \mu_{t,k,j}, s_j) = \mu_{t,k,j} + \mu^2_{t,k,j} s_j$. The model allows different dispersion levels across OTUs through OTU-specific overdispersion parameters s_j. In the limit as $s_j \rightarrow 0$, the model in Equation (1) yields the Poisson distribution with mean $\mu_{t,k,j}$. We assume a gamma distribution for a prior distribution of s_j, $s_j \overset{iid}{\sim} \text{Ga}(a_s, b_s), j = 1, \ldots, J$, with fixed a_s and b_s.

2.1.2. Model for Regression

We next model the mean count $\mu_{t,k,j}$ of $y_{t,k,j}$. We decompose the mean count into factors, a baseline mean count and a function of covariates, $\mu_{t,k,j} = g_{t,k,j}\eta_j(\boldsymbol{X}_t)$. Here parameter $g_{t,k,j}$ denotes the baseline mean abundance of OTU j in sample (t,k) and $\eta_j(\boldsymbol{X}_t)$ is a function of covariates $\boldsymbol{X}_t$ for OTU j to model the covariate effects. We construct a generalized regression model by letting $\log(\eta_j(\boldsymbol{X}_t)) = \boldsymbol{X}_t'\beta_j$, where $\beta_j = (\beta_{j1}, \ldots, \beta_{jP})'$ is a P-dimensional vector of regression coefficients of OTU j (Lawless, 1987; McCullagh and Nelder, 1989). The coefficient $\beta_{j,p}$ quantifies the effect of covariate p X_p on the mean abundance of OTU j. A vector β_j close to the zero vector produces a value of $\eta_j(\boldsymbol{X}_t)$ close to 1, and the mean count remains similar to the baseline mean count $g_{t,k,j}$, implying insignificant covariate effects. A negative (positive) of $\beta_{j,p}$ implies a negative (positive) association between mean counts and the p-th covariate, and a larger value of $X_{j,p}$ decreases (increases) the mean count, while holding the other covariates constant. We consider a Laplace prior on β_j. Specifically, we express the Laplace distribution as a scale mixture of normals and assume for $j = 1, \ldots, J$ and $p = 1, \ldots, P$,

$$\begin{aligned} \beta_{j,p} \mid \sigma_j^2, \phi_{j,p} &\overset{indep}{\sim} \mathrm{N}(0, \sigma_j^2\phi_{j,p}), \quad \phi_{j,p} \overset{indep}{\sim} \mathrm{Exp}(\frac{\lambda_j^2}{2}), \\ \lambda_j^2 &\overset{iid}{\sim} \mathrm{Ga}(a_\lambda, b_\lambda), \quad \sigma_j^2 \overset{iid}{\sim} \mathrm{IG}(a_\sigma, b_\sigma), \end{aligned} \tag{2}$$

where a_λ, b_λ, a_σ, and b_σ are fixed. σ_j^2 and $\phi_{j,p}$ denote the global and local shrinkage parameters, respectively, for OTU j. After integrating $\phi_{j,p}$ out, the prior distribution of $\beta_{j,p}$ is the Laplace distribution with location parameter 0 and scale parameter $\sqrt{\sigma_j^2}/\lambda_j$, that is, $p(\beta_{j,p} \mid \lambda_j^2, \sigma_j^2) \propto \exp\left(-\lambda_j|\beta_{j,p}|/\sqrt{\sigma_j^2}\right)$. Compared to a normal distribution that is a common choice for the prior of $\beta_{j,p}$, the Laplace distribution has more concentration around zero but allows heavier tails. The regularized regression through the Laplace prior more shrinks the coefficients of insignificantly related covariates into zero and less pulls the coefficients of important covariates toward zero. Shrinkage of β estimates through the model in Equation (2) mitigates possible issues due to multicollinearity and efficiently improves estimation of β in a high dimensional setting (Polson and Scott, 2012).

2.1.3. Model for Baseline Mean Count

We next build a prior probability model for the baseline mean count $g_{t,k,j}$ of OTU j in sample (t,k). We assume $g_{t,k,j} = r_{t,k}\alpha_{0,j}\alpha_{t,j}$ to separate sample $(r_{t,k})$, OTU $(\alpha_{0,j})$, and OTU-time $(\alpha_{t,j})$ factors. Sample total counts $y_{t,k,\cdot} = \sum_{j=1}^{J} y_{t,k,j}$ may greatly differ for different samples possibly due to experimental artifacts. For example, counts of an OTU even in the replicates taken at a time point may vastly differ. Sample specific size factors $r_{t,k}$ account for different total counts in different samples and expected counts normalized by $r_{t,k}$ are comparable across samples. Factor $\alpha_{0,j}$ explains variabilities in baseline mean abundances of OTUs and $\alpha_{t,j}$ models temporal dependence of the mean counts for an OTU, respectively. Factors $\alpha_{0,j}$ and $\alpha_{t,j}$ are not indexed by replicate k and account for stochastic change over time in normalized baseline expected counts of OTU j. Collecting all, we write the mean count as

$$\mu_{t,k,j} = g_{t,k,j}\eta_j(\boldsymbol{X}_t) = r_{t,k}\alpha_{0,j}\alpha_{t,j}\eta_j(\boldsymbol{X}_t), \tag{3}$$

The model for $g_{t,k,j}$ in Equation (3) is overparameterized and the individual parameters are not identifiable. To avoid potential identifiability issues, many of NB models rely on some form of approximation for the baseline mean counts. For example, one may find the maximum likelihood estimates (MLEs) of baseline mean abundance under some constraints and plug in those estimates to infer the mean abundance levels $\mu_{t_i,j}$ of OTUs (Witten, 2011). Plugging in MLEs is simple but may not be robust. In particular, the inference is greatly affected by a small change in a few OTUs that have large counts. Moreover, the errors introduced in the baseline mean count estimation will not be reflected in the inference. Several approaches to robustify the estimates are proposed (for example, see Anders and Huber, 2010; Witten, 2011). To circumvent the identifiability issue and provide uncertainty quantification for estimation of $g_{t,k,j}$, we take an alternative in Li et al. (2017) by imposing regularizing priors with mean constraints for $r_{t,k}$ and $\alpha_{0,j}$. We let the logarithm of the factors $\tilde{r}_{t,k} = \log(r_{t,k})$ and $\tilde{\alpha}_{0,j} = \log(\alpha_{0,j})$, and assume the regularizing prior distribution with mean constraints,

$$\begin{aligned} \tilde{r}_{t_i,k} \mid \psi^r, \eta^r, w^r, \nu_r^2, c_r &\overset{iid}{\sim} \sum_{\ell=1}^{L^r} \psi_\ell^r \left\{ w_\ell^r \phi(\eta_\ell^r, \nu_r^2) + (1 - w_\ell^r)\phi\left(\frac{c_r - w_\ell^r\eta_\ell^r}{1 - w_\ell^r}, \nu_r^2\right)\right\}, \\ \tilde{\alpha}_{0,j} \mid \psi^\alpha, \eta^\alpha, w^\alpha, \nu_\alpha^2, c_\alpha &\overset{iid}{\sim} \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha \left\{ w_\ell^\alpha \phi(\eta_\ell^\alpha, \nu_\alpha^2) + (1 - w_\ell^\alpha)\phi\left(\frac{c_\alpha - w_\ell^\alpha\eta_\ell^\alpha}{1 - w_\ell^\alpha}, \nu_\alpha^2\right)\right\}, \end{aligned} \tag{4}$$

where $\phi(\eta, \nu^2)$ is the probability density function of the normal distribution with mean η and variance ν^2, constraints for the mixture weights $\sum_{\ell=1}^{L^r} \psi_\ell^r = \sum_{\ell=1}^{L^\alpha} \psi_\ell^\alpha = 1$ with $0 < \psi_\ell^r < 1$ and $0 < \psi_\ell^\alpha < 1$, $0 < w_\ell^r < 1$, and $0 < w_\ell^\alpha < 1$ for all ℓ. Mixture models as in Equation (4) are often used as a basis to approximate any distribution. Each component in Equation (4) is further composed of a mixture of two normals, $\mathrm{N}(\eta_\ell, \nu^2)$ and $\mathrm{N}\left(\frac{(c - w_\ell\eta_\ell)}{(1 - w_\ell)}, \nu^2\right)$ with weights w_ℓ and $1 - w_\ell$, respectively, and the mean of the component is c. In consequence, the prior and posterior of $\tilde{r}$ and $\tilde{\alpha}$ under the model in Equation (4) satisfy their prespecified mean constraints c_r and c_α, respectively. Li et al. (2017) showed that the model in Equation (4) flexibly accommodates various features in a distribution such as skeweness or multi-modality while satisfying the constraints. Furthermore, the model based normalization through Equation (4) enables joint analysis of all OTUs and can further improve estimation of the covariate effects. With the regularizing priors, baseline mean counts $g_{t,k,j}$ are identifiable, while $r_{t,k}$, $\alpha_{0,j}$, and $\alpha_{t,j}$ are not directly interpretable. More importantly, the parameters of primary interest $\eta_j(\boldsymbol{X}_t)$ can be uniquely estimated and $\beta_{j,p}$'s keep their interpretation as parameters that quantify the effects of covariates on mean counts. We used an empirical approach to fix the mean constraints c_r and c_α. Sensitivity analyses were conducted to assess the robustness to the specification of c_r and c_α and show that the model provides reasonable estimates of $g_{t,k,j}$ and moderate changes in the values of c_r and c_α minimally change the estimates. More details of the specification of c_r and c_α are discussed in section 3.1. We fix the numbers of mixture components, L^r and L^α and variances v_r^2 and v_α^2. We let $\eta_\ell^r \overset{iid}{\sim} \mathrm{N}(c_r, \omega_r^2)$ and $\eta_\ell^\alpha \overset{iid}{\sim} \mathrm{N}(c_\alpha, \omega_\alpha^2)$, where ω_r^2 and ω_α^2 are fixed. We assume $\psi^r = (\psi_1^r, \ldots, \psi_{L^r}^r) \sim \mathrm{Dir}(d_r, \ldots, d_r)$ and $\psi^\alpha = (\psi_1^\alpha, \ldots, \psi_{L^\alpha}^\alpha) \sim \mathrm{Dir}(d_\alpha, \ldots, d_\alpha)$, with fixed d_r and d_α. We let $w_\ell^r \overset{iid}{\sim} \mathrm{Be}(a_r, b_r)$, $\ell = 1, \ldots, L^r$ and $w_\ell^\alpha \overset{iid}{\sim} \mathrm{Be}(a_\alpha, b_\alpha)$, $\ell = 1, \ldots, L^\alpha$ with fixed a_r, b_r, a_α, and b_α.

Recall that samples are collected over time points $t_1, \ldots, t_n$ in $[0, T]$ and $\alpha_{t,j}$ accounts for temporal dependence in the baseline mean count for an OTU. We let $\tilde{\alpha}_{t,j} = \log(\alpha_{t,j})$ a function in time t and use a stochastic process to model temporal dependence among $\mu_{t,k,j}$. The Gaussian process (GP) is one of the most popular stochastic models for the underlying process in spatial and spatio-temporal data (for example, see Cressie, 1992; Banerjee et al., 2014 among many others). The GP effectively represents the underlying phenomenon in a variety of applications, but it has some drawbacks such as a complex computation that requires a matrix decomposition and problematic estimation of the parameters in its covariance function, potentially leading to difficulties in exploring the posterior distribution (Lee et al., 2005; Liang and Lee, 2014). To alleviate such difficulties of GP models while still maintaining their flexibility and adaptiveness, we use a convolution approach with a kernel function developed in Higdon (1998, 2002). For each OTU, we specify the latent process $\boldsymbol{\theta}_j(t)$ to be nonzero only at the time points $u_1, \ldots, u_M$ in $[0, T]$. Specifically, we consider the GP convolution model,

$$\tilde{\alpha}_{t,j} = \sum_{m=1}^{M} Z(t - u_m)\theta_{m,j},$$

where $\{u_1, \ldots, u_M\}$ a set of basis points in $[-t_1', T + t_2']$ with $t_1', t_2' > 0$, and $Z(t - u_m)$ a Gaussian kernel centered at u_m, $Z(t - u_m) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\{\frac{-(t-u_m)^2}{2\gamma^2}\}$. The number of basis points M, their locations u_m and the range parameter γ can be treated as random variables by placing prior distributions, e.g., consider a gamma prior for γ. For simplicity, we fix them as follows. We first choose a value for M and let u_m evenly spaced over time $[-t_1', T + t_2']$. Following Xiao (2015), we let the range parameter $\gamma^2 = ((2T + t_1' + t_2')/M)^2$, that is, the range parameter depends on the value of M. Through simulations, we studied the impact of different values of M on the posterior inference of $g_{t,k,j}$. A discussion is included in section 3.1. Given the number of basis points M, we assume $\theta_{m,j} \mid \tau_j^2 \overset{indep}{\sim} \mathrm{N}(0, \tau_j^2)$ and $\tau_j^2 \overset{iid}{\sim} \mathrm{IG}(a_\tau, b_\tau)$, $m = 1, \ldots, M$ and $j = 1, \ldots, J$.

We implement posterior inference on the parameters $\tilde{\boldsymbol{\theta}} = (\beta_j, \sigma_j^2, \lambda_j^2, \phi_{j,p}, \tilde{r}_{t,k}, \psi^r, w_\ell^r, \eta_\ell^r, \tilde{\alpha}_{0,j},\ \tilde{\alpha}_{t,j}, \psi^\alpha, w_\ell^\alpha, \eta_\ell^\alpha, \boldsymbol{\theta}_j, \tau_j^2, s_j)$ via a Markov chain Monte Carlo (MCMC) method based on Metropolis-Hastings algorithm and Gibbs sampling. Each of the parameters is iteratively updated conditional on the currently computed values of all other parameters to simulate a sample from the posterior distribution. The parameters $\tilde{r}$ and $\tilde{\alpha}_0$ jointly determine baseline mean counts and joint updating of $\tilde{r}$ and $\tilde{\alpha}_0$ may greatly improve the mixing. In our ocean microbiome data, some discretized covariates are missing. We treat them as random variables by assuming a uniform distribution over possible categories, and impute their values in MCMC simulation. Full details of our MCMC algorithm are given in Supplementary section 1. We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots and autocorrelation plots of imputed parameters. For the upcoming simulation examples and the data analysis, we found no evidence of practical convergence problems. An R package of the code used for simulations and the analysis of the ocean microbiome dataset in the following sections is available from the authors website https://users.soe.ucsc.edu/~juheelee/.

2.2. Simulation Experiment: Data Generation and Comparative Study

We conducted simulation studies to assess the performance of our model. We compared the model to an alternative model, the negative binomial mixed model (NBMM) in Zhang et al. (2017). We assumed a sample of $J = 200$ OTUs. We used the same time points (t_i) and numbers of replicates (K_i) of our ocean microbiome data as shown in **Figure 1B**. We let $\beta_{j,p}^{\mathrm{TR}} = 0$ with probability 0.85. For $\beta_{j,p}^{\mathrm{TR}} \neq 0$ we simulated $\beta_{j,p}^{\mathrm{TR}}$ from either of $\mathrm{N}(-1.5, 0.05^2)$ or $\mathrm{N}(1.5, 0.05^2)$ with equal probability, where $\mathrm{N}(a, b^2)$ denotes the normal distribution with mean a and variance b^2. It implies that a covariate has no effect on OTU abundance with probability 0.85 or may significantly affect mean abundance with the remaining probability 0.15. To specify $r_{t,k}^{\mathrm{TR}}$ and $\alpha_{0,j}^{\mathrm{TR}}$, we did not assume any distribution and used their classical estimates from our ocean microbiome data; following Witten (2011), we first computed estimates of sample size factors $r_{t_i,k}'$ and OTU size factors $\alpha_{0,j}'$ using the ocean microbiome data, $r_{t_i,k}' = y_{t_i,k,\cdot}/y_{\ldots}$ and $\alpha_{0,j}' = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K_i} y_{t_i,k,j}/r_{t_i,k}'$ where $y_{t_i,k,\cdot} = \sum_{j=1}^{J} y_{t_i,k,j}$ and $y_{\ldots} = \sum_{j=1}^{J} y_{\cdot\cdot,j}$. We then randomly sampled from the pool of $r_{t,k}'$ and $\alpha_{0,j}'$ to specify the true values. To simulate temporal dependence in OTU abundance, we let $\tilde{\alpha}_{t_i,j}^{\mathrm{TR}} = a_{t_i,j}\cos(2\pi(\tilde{t}_i - b_{t_i,j})) + c_{t_i,j}(\tilde{t}_i - \tilde{t}^\star)^2$. Here $\tilde{t}_i$ denotes time t_i in year and $\tilde{t}^\star$ the median of $\tilde{t}_i$. We let $a_{t,j} \overset{iid}{\sim} \mathrm{N}(0.15, 0.1^2)$, $b_{t,j} \overset{iid}{\sim} \mathrm{N}(0, 0.5^2)$, and $c_{t,j} \overset{iid}{\sim} \mathrm{N}(0.1, 0.1^2)$ to have different patterns for OTUs. For some OTUs, $\tilde{\alpha}_{t_i,j}^{\mathrm{TR}}$ are illustrated in red squares in **Figures 4E–G**. We generated $s_j^{\mathrm{TR}} \overset{iid}{\sim} \mathrm{Ga}(1, 10)$. We used the covariate matrix of the ocean microbiome data illustrated in **Figure 2** for the simulation study. For the missing covariates in the data, we generated a value of possible categories with equal probability. We finally simulated OTU counts $y_{t_i,k,j}$ from the negative binomial distribution $y_{t_i,k,j} \overset{indep}{\sim} \text{NB}(\mu^{\text{TR}}_{t_i,k,j}, s_j^{\text{TR}})$, where $\mu^{\text{TR}}_{t_i,k,j} = r^{\text{TR}}_{t_i,k}\alpha^{\text{TR}}_{0,j} \exp(\tilde{\alpha}^{\text{TR}}_{t_i,j} + \boldsymbol{X}^{\text{TR}}_t \beta^{\text{TR}}_j)$.

For comparison, we used the negative binomial mixed model (NBMM) in Zhang et al. (2017). Similar to the proposed model, the NBMM uses a negative binomial distribution with mean μ^{NBMM} and shape parameter θ^{NBMM} to model OTU counts and assumes $\log(\mu^{\text{NBMM}}_{t,k,j}) = \log(y_{t,k,\cdot}) + \beta^{\text{NBMM}}_{0,j} + \boldsymbol{X}_t \beta^{\text{NBMM}}_j + \boldsymbol{Z}_{t,k} \boldsymbol{b}^{\text{NBMM}}_j$ where $\boldsymbol{X}_t$ and $\boldsymbol{Z}_{t,k}$ are the covariate matrices for fixed effects and random effects, respectively. It assumes random effects $\boldsymbol{b}^{\text{NBMM}}_j \sim \text{N}(\boldsymbol{0}, \Psi)$. By letting the replicates at a time point share the same random effect, OTU abundances in the replicates at a time point are correlated. The NBMM normalizes OTU counts by sample total counts. That is, sample total counts $y_{t,k,\cdot}$ are used as an offset to adjust for the variability in total counts across samples. Similar to other existing methods, the NBMM performs separate analyses of OTUs. An iterative weighted least squares algorithm is developed to produce the MLEs under the NBMM and implemented in a R function *glmm* in R package *BhGLM*.

2.3. Ocean Microbiome Data: Data Description and Preprocessing

We applied the proposed statistical method to ocean microbiome data. Seawater samples were collected weekly at the end of Santa Cruz Municipal Wharf (SCW), Monterey Bay (36.958 °N, −122.017 °W), with an approximate depth of 10 m. SCW is one of the ocean observing sites in Central and Northern California (CenCOOS), where harmful algal bloom species [HAB species: *Alexandrium* (Ax), *Dinophysis* (Dp), *Pseudo-nitzschia* (Pn) etc.] are monitored weekly along with nutrient measurements [ammonia (NH_4), silicate (Si), nitrate (N), phosphate (P)], temperature (T), domoic acid (DA) concentration, and chlorophyll (Chl). Details of phytoplankton net tow sampling of measuring phytoplankton abundance, measurement of physical (nutrients and temperature) and biological parameters (chlorophyll α and DA concentration) are described in Sison-Mangus et al. (2016). *Pseudo-nitzschia*, *Dinophysis*, and *Alexandrium* cells were counted with a Sedgewick rafter counter under the microscope. Data for physical and biological factors are available from the website link http://www.sccoos.org/query/?project=Harmful%20Algal%20Blooms&study[]=Santa%20Cruz%20Wharf. Among the 10 variables, the concentration levels of *Alexandrium*, *Dinophysis*, *Pseudo-nitzschia*, and domoic acid have highly right-skewed distributions and are discretized into categories based on their biological properties for our analysis. The ranges of the concentration levels for the discretization are in Supplementary Table 1 and **Figures 2A–J** illustrates all covariates included for analysis. The values of −1, 0, 1, 2, 3, and 4 represent missing values and the categories of None, Low, Medium, High, and Very High for the discretized variables, respectively. Due to high right skeweness, categories corresponding to high concentration levels have low frequencies. Values of the *Dinophysis* concentration level are missing at 20 time points among 55 points used for analysis. Sample correlations between the factors are relatively strong. **Figures 2K,L** shows scatterplots for some selected pairs of the factors.

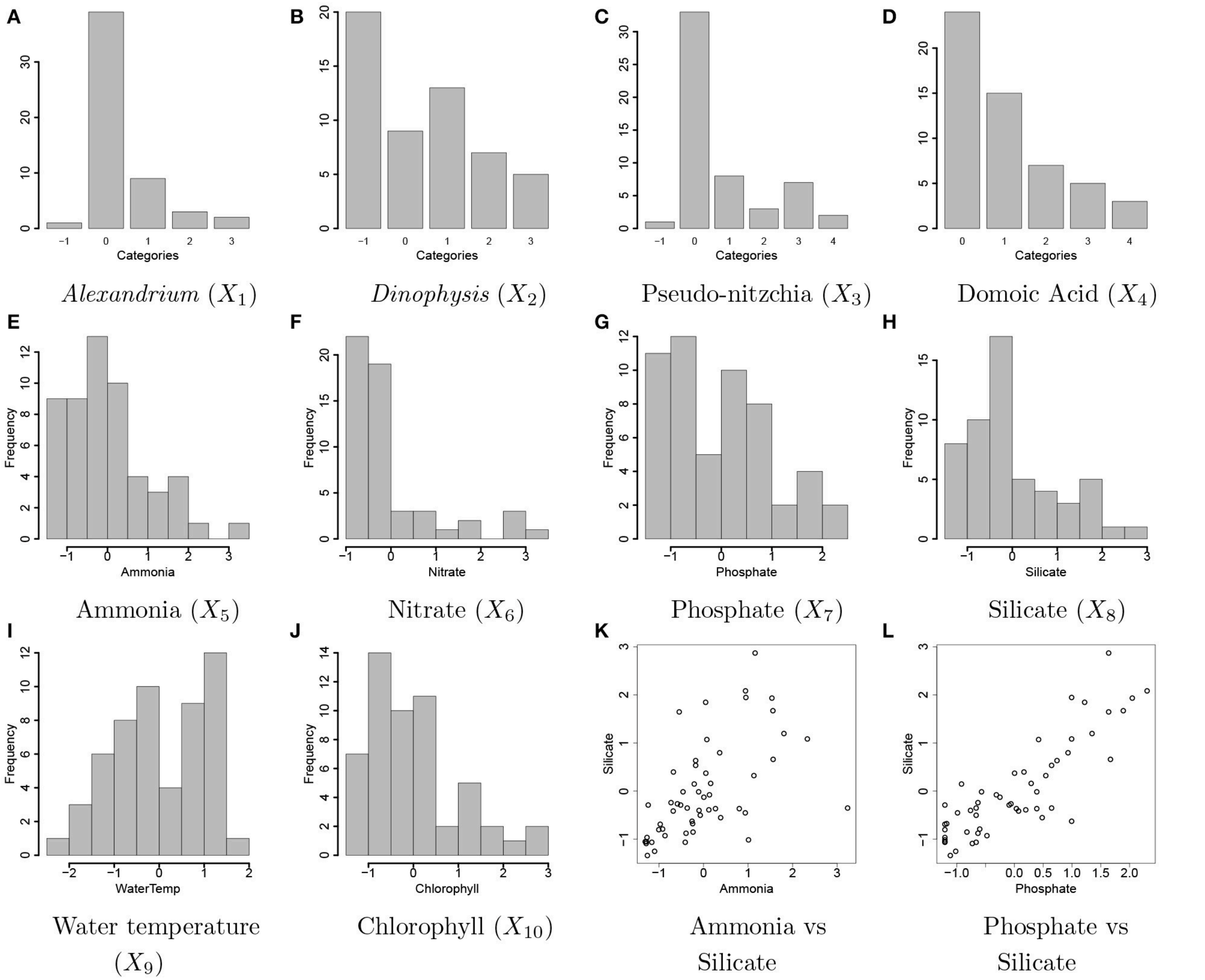
For bacterial RNA samples, three depth-integrated (0, 5, and 10 ft) water samples were collected at a total of 55 time points between April 2014 and November 2015. Two or three samples are sequenced at each time point. The numbers of replicates at the time points are illustrated in **Figure 1B**. Microbial RNA in the samples was extracted for 16S rRNA sequencing. Post-processing of sequences was performed using the Quantitative Insights Into Microbial Ecology (QIIME 1.9.1) pipeline. A total of nearly 39,823 OTUs were obtained in data after removing singletons. We restricted our attention to OTUs that have greater than or equal to five counts on average. The rule leaves in the end $J = 263$ OTUs for the 150 samples for the analysis. A heatmap of the counts in the filtered data is shown in **Figure 1**. The primary goal of the study is to investigate the effects of physical and biological factors on abundance levels of OTUs, while accounting for baseline abundance levels of OTUs in samples.

3. RESULTS

3.1. Simulation Experiment: Model Fitting and Comparison

To fit the proposed model for the simulated data designed in section 2.2, we specified hyperparameter values of the model as follows; for the Laplace prior of $\beta_{j,p}$, we let $a_\lambda = b_\lambda = 0.5$ for a gamma prior of λ_j^2 (with mean a_λ/b_λ and variance a_λ/b_λ^2) and $a_\sigma = b_\sigma = 0.3$ for an inverse gamma prior for common variances σ_j^2. For the regularizing priors of $\tilde{r}_{t_i,k}$ and $\tilde{\alpha}_{0,j}$, we fixed $d_\alpha = d_r = 10$, $a_r = b_r = a_\alpha = b_\alpha = 1$, $\omega_r^2 = \omega_\alpha^2 = 1.0$, $v_r^2 = 1$, and $v_\alpha^2 = 2.0$. We also fixed the number of mixture components for the regularizing priors $L_r = 30$ and $L_\alpha = 50$. To specify values of the mean constraints c_r and c_α, we took an empirical approach. We used the simulated $y_{t_i,k,j}$, computed estimates of $r_{t_i,k,j}$ and $\alpha_{j,0}$ as described in section 2.2 and fixed the mean constraints at the means of the logarithm of the estimates, respectively. Note that the specified values of c_r and c_α were very different from the means of their true values. For the process convolution prior of OTU-time factor $\tilde{\alpha}_{t_i,j}$, we chose a value of M such that the kernel function at a basis point is not entirely located in a place where no sample is obtained. We let the number of basis $M = 13$ and basis u_m, $m = 1, \ldots, M$ evenly spaced between −10 and $T_i + 10$. For overdispersion parameter s_j we let $a_s = 1$ and $b_s = 2$. To run MCMC simulation, we initialized the parameters by simulating with their prior distributions. We then implemented posterior inference using MCMC simulation over 25,000 iterations, discarding the first 10,000 iterations as burn-in and choosing every other sample as thinning.

Figure 3 illustrates the comparison of posterior estimates of $\beta_{j,p}$ to their true values $\beta^{\text{TR}}_{j,p}$ for some selected covariates. In the figure, dots and blue dashed lines represent posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$ and their 95% credible intervals, respectively. $\hat{\beta}_{j,p}$s are around the 45 degree line (red dotted line) for most of (j, p) and most of the interval estimates captures the true values. It implies that the proposed model reasonably recovers $\beta^{\text{TR}}_{j,p}$. For categories 3 and 4 of X_4 in **Figures 3I,J**, the credible intervals tend to be wider due to their low frequencies in the data as shown



Alexandrium (X_1) *Dinophysis* (X_2) Pseudo-nitzchia (X_3) Domoic Acid (X_4)

Ammonia (X_5) Nitrate (X_6) Phosphate (X_7) Silicate (X_8)

Water temperature (X_9) Chlorophyll (X_{10}) Ammonia vs Silicate Phosphate vs Silicate

FIGURE 2 | Ocean microbiome data. Bar plots of discretized covariates, concentration levels of *Alexandrium* (Ax, X_1) and *Dinophysis* (Dp, X_2), Pseudo-nitzchia (Pn, X_3), domoic acid (DA, X_4) in **(A–D)**. The values of −1, 0, 1, 2, 3, and 4 represent a missing value, none and low, medium, high, and highest concentration levels, respectively. Histograms of continuous covariates, concentration levels of ammonia (NH_4, X_5), nitrate (N, X_6), phosphate (P, X_7), silicate (Si, X_8), water temperature (T, X_9), concentration level of chlorophyll (Chl, X_{10}) in **(E–J)**. The variables are standardized to have mean 0 and variance 1 prior to analysis. A scatterplot of the concentrations of ammonia and silicate and a scatterplot of the concentrations of phosphate and silicate are shown in **(K,L)**, respectively.

in **Figure 2D**. The insert plot in each panel illustrates a scatter plot of $\hat{\beta}_{j,p}$ for (j,p) with $\beta^{\mathrm{TR}}_{j,p} = 0$. It shows that the proposed regression model with the Laplace prior effectively shrinks $\beta_{j,p}$ with $\beta^{\mathrm{TR}}_{j,p} = 0$ to zero, as is desired in our simulation setup. Supplementary Figure 1 has plots for all covariates.

Figures 4A–C illustrate plots of $g^{\mathrm{TR}}_{t,k,j}$ vs estimates of $g_{t,k,j}$ with their means (black dots) and 95% credible intervals (blue vertical lines) for some selected OTUs, $j = 8, 34$, and 48. Recall that we do not attempt to recover the true values of individual $r_{t_i,k}$, $\alpha_{0,j}$, and $\alpha_{t,j}$, but we rather aim to reasonably recover the true baseline mean counts, $g^{\mathrm{TR}}_{t_i,k,j} = r^{\mathrm{TR}}_{t_i,k}\alpha^{\mathrm{TR}}_{0,j}\exp(\tilde{\alpha}^{\mathrm{TR}}_{t_i,j})$. In the figure the estimates are tightly around the 45 degree line, providing evidence that reasonable estimates of baseline mean counts are obtained under the proposed model. **Figure 4D** has a histogram of averaged differences between baseline mean count estimates and their true values, $D_j = \sum_{i=1}^{n}\sum_{k=1}^{K_i}(\hat{g}_{t_i,k,j} - g^{\mathrm{TR}}_{t_i,k,j})/N$. The averaged differences are around zero, implying that the proposed model provides reasonable estimates of baseline mean counts for most of OTUs. We further examined individual parameters. **Figures 4E–G** shows the comparison of estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ to their true values over time for the same OTUs in **Figures 4A–C**. Black dots and blue vertical lines represent estimates of posterior means of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ and their 95% credible intervals, respectively. Red squares represent their true values. From the figure, the estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ are consistently greater than their true values at all time points, but capture their overall temporal trend. **Figure 4H** illustrates a scatterplot of $\tilde{r}^{\mathrm{TR}}_{t,k}$ and their posterior estimates of $\tilde{r}_{t,k}$, where dots and blue vertical

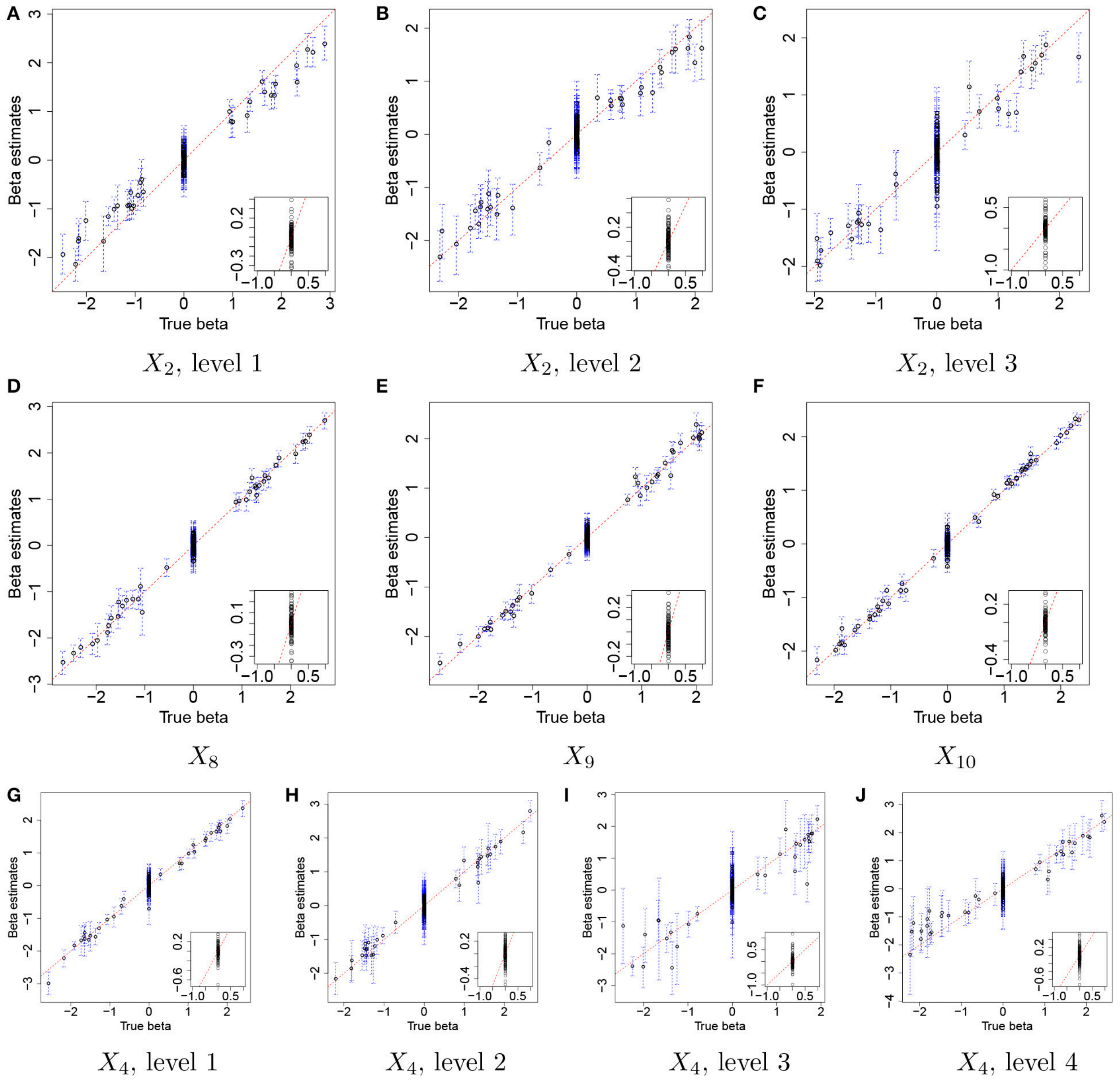
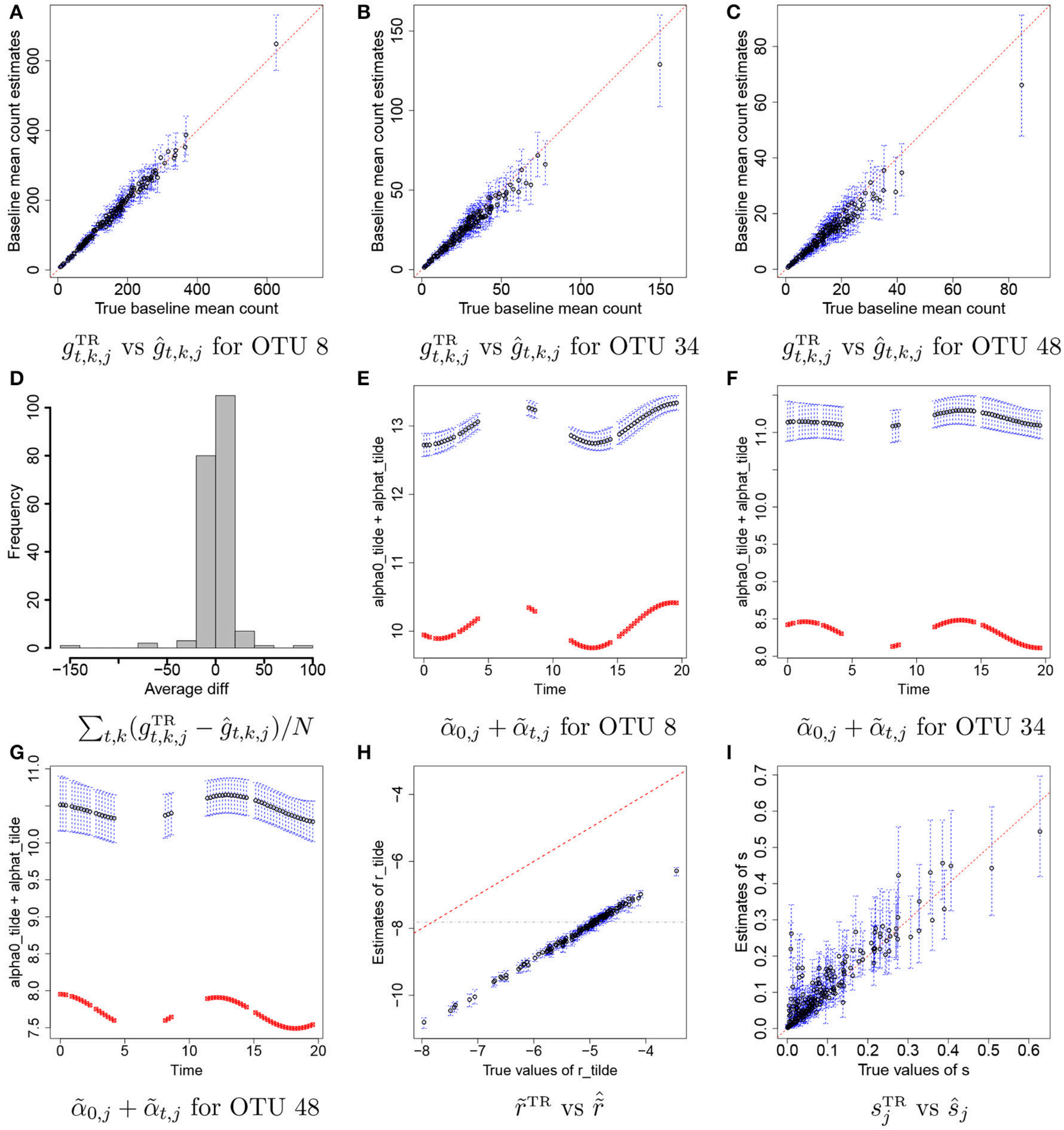


FIGURE 3 | Simulation 1—proposed model. Comparison of the true values $\beta_{j,p}^{\text{TR}}$ and the posterior estimates of $\beta_{j,p}$ under the proposed model for some selected covariates. Dots and blue dashed lines represent estimates of posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$ and 95% credible intervals (CIs) of $\beta_{j,p}$, respectively. The insert plot in each panel is a scatter plot of $\hat{\beta}_{j,p}$ and $\beta_{j,p}^{\text{TR}}$ for (j, p) with $\beta_{j,p}^{\text{TR}} = 0$.

intervals denote estimates of posterior means and 95% credible intervals, respectively, and the gray horizontal line is at c_r used for analysis. Different from the estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$, the estimates of $\tilde{r}_{t,k}$ fall below the 45 degree line approximately by the same distance for all OTUs. It shows that estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ and $\tilde{r}_{t,k}$ have discrepancies from their true values but in the opposite direction and the model can produce reasonable estimates of $g_{t,k,j}$ as seen in **Figures 4A–D**. The true overdispersion parameters s_j^{TR} are reasonably well estimated as shown in **Figure 4I**. We check the posterior predictive distribution of $Y_{t,k,j}$. The posterior predicted values of $Y_{t_i,k,j}$ with their 95% predictive intervals for OTUs $j = 8, 34$, and 48 are compared to their observed values in Supplementary Figure 2. The figure indicates a reasonable model fit.

In addition, we conducted a sensitivity analysis to the specification of mean constraints c_r and c_α for the priors of $\tilde{r}$ and $\tilde{\alpha}_0$. We used different values for c_r and c_α and compared the estimates of $g_{t,k,j}$ to their truth. Supplementary Figures 3a–c has histograms of averaged differences D_j between $\hat{g}_{t_i,k,j}$ and $g_{t_i,k,j}^{\text{TR}}$ for different specification of c_r and c_α. The histograms show



$g^{\mathrm{TR}}_{t,k,j}$ vs $\hat{g}_{t,k,j}$ for OTU 8 $g^{\mathrm{TR}}_{t,k,j}$ vs $\hat{g}_{t,k,j}$ for OTU 34 $g^{\mathrm{TR}}_{t,k,j}$ vs $\hat{g}_{t,k,j}$ for OTU 48

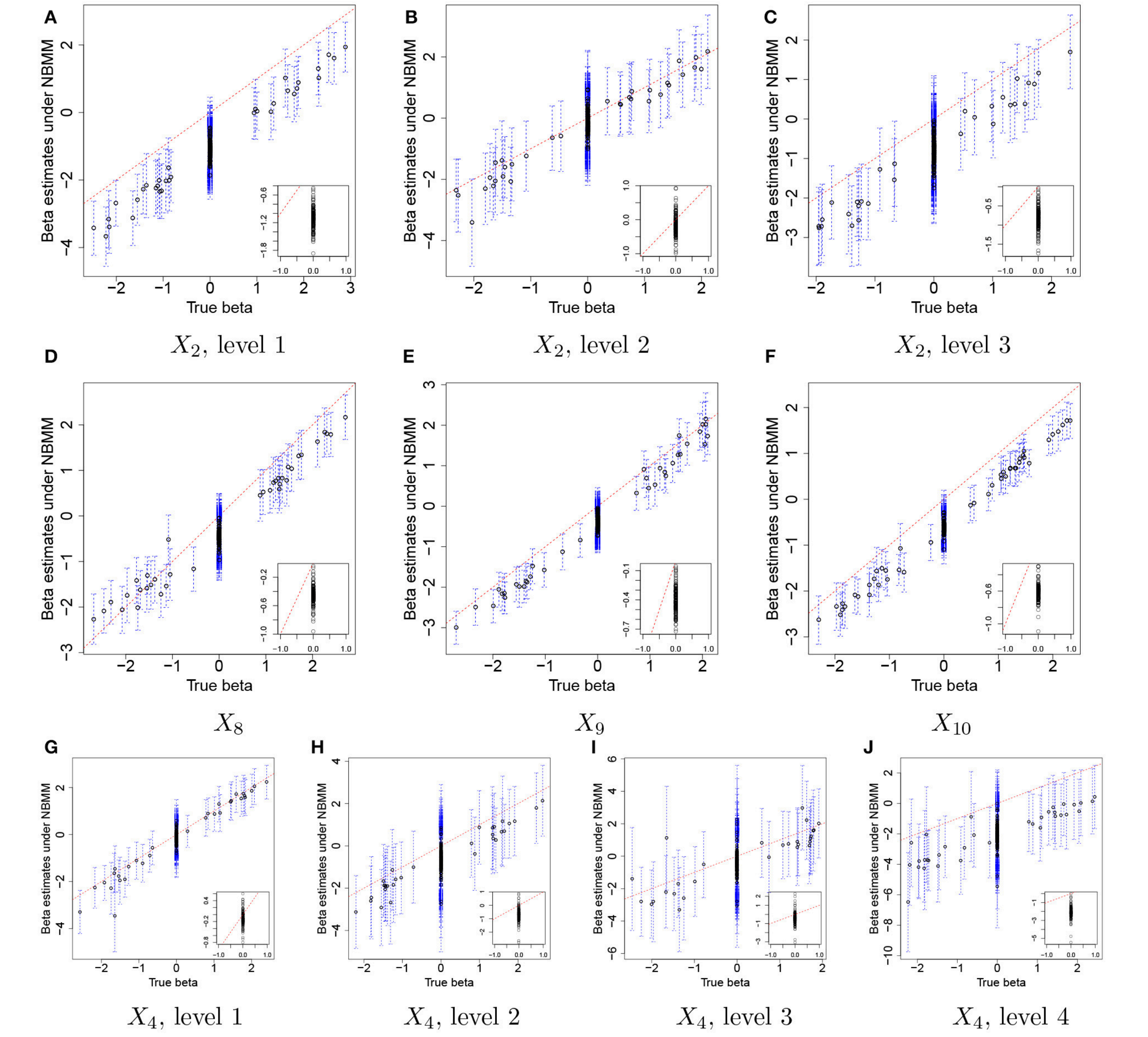
$\sum_{t,k}(g^{\mathrm{TR}}_{t,k,j} - \hat{g}_{t,k,j})/N$ $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ for OTU 8 $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ for OTU 34

$\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ for OTU 48 $\tilde{r}^{\mathrm{TR}}$ vs $\hat{\tilde{r}}$ s^{TR}_j vs $\hat{s}_j$

FIGURE 4 | Simulation 1—proposed model. Panels **(A–C)** illustrate plots of the true baseline mean counts $g^{\mathrm{TR}}_{t,k,j}$ vs their estimates $\hat{g}_{t,k,j}$ for some selected OTUs $j = 8, 34, 48$. Panel **(D)** shows a histogram of averaged differences between $g^{\mathrm{TR}}_{t,k,j}$ and $\hat{g}_{t,k,j}$ for each OTU. Panels **(E–G)** show plots of estimates of $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ over time for OTUs $j = 8, 34, 48$. Panel **(H)** has a scatterplot of $\hat{\tilde{r}}_{t,k}$ vs. $\tilde{r}^{\mathrm{TR}}$. Panel **(I)** has a scatterplot of $\hat{s}$ vs. s^{TR}. Dots represent posterior mean estimates and blue vertical dotted lines 95% credible intervals. Red squares represent the true values.

minor change in estimates of $g_{t_i,k,j}$ under different specifications of c_r and c_α. An sensitivity analysis to the specification of the number M of basis points in the GP convolution model for $\tilde{\alpha}_{t,j}$ was also performed. We used $M = 8$, 13, and 18 and examined estimates of the baseline mean counts, $g_{t,k,j}$. Supplementary Figures 3a,d,e has histograms of averaged differences D_j for each of M. The results indicate that the baseline mean counts are reasonably estimated for a range of values of M in the simulation study.

For comparison, we used the NBMM to the simulated data. Since the NBMM does not accommodate missing covariates, we used $\boldsymbol{X}^{\mathrm{TR}}$ to fit the NBMM. **Figure 5** compares the MLEs $\hat{\beta}^{\mathrm{NBMM}}_{j,p}$ of $\beta_{j,p}$ to the true values for the same covariates used in **Figure 3**. Dots and blue vertical lines represent the MLEs under the NBMM and their 95% confidence intervals, respectively. Comparing **Figure 5** to **Figure 3**, the NBMM produces poor estimates. The MLEs are biased for some covariates (e.g., **Figure 5A**). Also, confidence intervals under the NBMM often



X_2, level 1 X_2, level 2 X_2, level 3

X_8 X_9 X_{10}

X_4, level 1 X_4, level 2 X_4, level 3 X_4, level 4

FIGURE 5 | Simulation 1—NBMM. Comparison of the true values $\beta^{\text{TR}}_{j,p}$ and maximum likelihood estimates $\hat{\beta}^{\text{NBMM}}_{j,p}$ of $\beta_{j,p}$ under the negative binomial mixed model (NBMM) for some selected covariates. Dots and blue dashed lines represent $\hat{\beta}^{\text{NBMM}}$ and their 95% confidence intervals, respectively. The insert plot in each panel is a scatter plot of $\hat{\beta}^{\text{NBMM}}_{j,p}$ and $\beta^{\text{TR}}_{j,p}$ for (j, p) with $\beta^{\text{TR}}_{j,p} = 0$.

fail to capture the true values and their interval estimates under the NBMM tend to be much wider than those under the proposed model. Normalization through observed sample total counts and inducing correlation in replicates through iid (independent and identically distributed) random effects under the NBMM may lead to poor estimation of the baseline mean abundance for the simulated data, resulting in deterioration of coefficient estimation. In addition, separate analyses of OTUs under the NBMM do not allow to strengthen estimates through combining information across OTUs. Comparing the insert plots in **Figure 5** to those in **Figure 3**, $\hat{\beta}^{\text{NBMM}}_{j,p}$ with $\beta^{\text{TR}}_{j,p} = 0$ tends to more widely spread out from zero and often their confidence intervals fail to capture zero. Supplementary Figure 4 has plots of $\beta_{j,p}$ for all covariates. Supplementary Figures 4, 5 shows the comparison of the estimates $\hat{\theta}^{\text{NBMM}}$ of overdispersion parameters under the NBMM to their true values. Note that θ^{NBMM} is the inverse of s in our model. The NBMM underestimates s_j for many OTUs, and yields poor predicted values, implying the lack of a fit.

We further examined the performance of the proposed model through additional simulation studies, Simulations 2 and 3 in

Supplementary section 2. In these simulations, we studied the robustness of the model when different simulation setups are used to simulate data. In Simulation 2, we assumed no temporal dependence among OTU abundance and generated independent samples from a normal distribution for $\tilde{\alpha}_{t,j}$. We assumed that all $\beta^{\mathrm{TR}}_{j,p}$ has nonzero effects for all OTUs and simulated $\beta_{j,p}$ from a mixture of normals. The performance of our model is almost the same as in Simulation 1 (see Supplementary Figures 6–8). Interestingly, the NBMM that assumes iid random effects performs poorly for β estimation. In Simulation 3, we simulated $\tilde{\alpha}^{\mathrm{TR}}_{t,j}$ from a discontinuous function. The results show that when the temporal dependence pattern is not smooth as assumed for the GP, estimates of the baseline mean counts under the proposed model are slightly deteriorated but the model produces reasonable inference on $\beta_{j,p}$ (see Supplementary Figures 9–11). A more detailed summary of the additional simulations is given in Supplementary section 2.

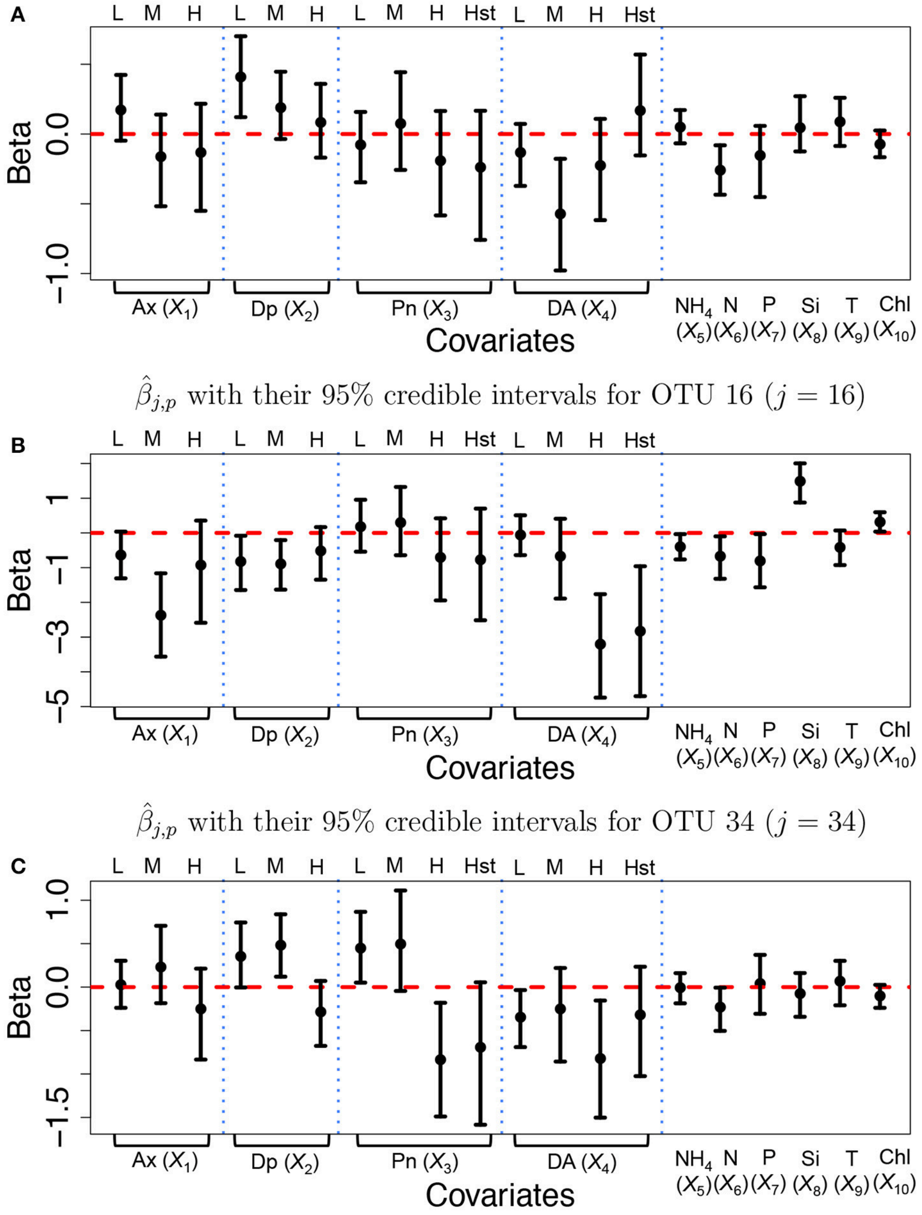
3.2. Ocean Microbiome Data: Model Fitting and Comparison

We specified hyperparameters similar to those in the simulations and analyzed the microbiome data in section 2.3. The MCMC simulation was run over 25,000 iterations. The first 15,000 iterations were discarded as burn-in and every other sample was kept as thinning and used for inference. **Figure 6** illustrates inference on covariate effects for some selected OTUs, $j = 16, 34$, and 49, taxonomically belonging to *Alteromonadales*, *Halomonas* sp., and *Alteromonadales* in the Gamma-proteobacteria phyla, respectively. Dots and vertical solid lines represent the posterior mean $\hat{\beta}_{j,p}$ and 95% credible interval estimates, respectively. Similar to the results of the simulation study, the credible intervals for high and highest levels of the discretized covariates tend to be wider due to their low frequencies in the data. From **Figure 6A**, on average the medium concentration level of domoic acid (DA, X_4) and the concentration level of nitrate (N, X_6) significantly decrease the mean abundance of OTU 16 by a multiplicative factor of $\exp(-0.572) = 0.564$ and $\exp(-0.260) = 0.771$, respectively. One may infer that the medium concentration level of domoic acid is significantly associated with lower expected counts for the OTU compared to those with category none of the domoic acid concentration level. A similar argument can be applied to the inference on the nitrate concentration level. Interestingly, we observed statistically significant reduction in abundance from many OTUs belonging to Gamma-proteobacteria including those OTUs for increasing domoic acid concentration (not shown). The resulting inference was further validated through a lab experiment. Most notably, four bacterial cultured isolates belonging to Gamma-proteobacteria (three among them are *Alteromonadales*) were observed to be severely retarded in growth after 2 days of exposure to increasing domoic acid of 0 to 150 μg/ml in the experiment (Sison-Mangus, unpublished data). This demonstrates that the proposed model successfully identifies important OTUs in ocean bacterial community dynamics for further investigation. More results are presented in Supplementary section 3. Supplementary Figures 12a–c illustrates the posterior estimates of baseline expected counts $\tilde{\alpha}_{0,j} + \tilde{\alpha}_{t,j}$ normalized by sample size factors for the OTUs. From the figure, the baseline expected counts vary over time for those OTUs and the temporal pattern of OTU $j = 34$ is different from those of OTUs $j = 16$ and 49. Histograms of the posterior mean estimates $\hat{\beta}_{j,p}$ of $\beta_{j,p}$, are illustrated in Supplementary Figure 13. The figure does not show clear overall tendency in the direction of association between covariates and OTU counts. Posterior inference on sample size factors $r_{t_i,k}$ and OTU specific overdispersion parameters s_j is illustrated in Supplementary Figures 12d,e.

For comparison, we fitted the NBMM to the data. Since the NBMM does not account for missing values, we use the maximum a posteriori estimates of the missing values under the proposed for the NBMM. We used the R function *glmm* and the algorithm produced warning messages on convergence for 32 OTUs. **Figure 7** illustrates $\hat{\beta}^{\mathrm{NBMM}}_{j,p}$ (dots) with their 95% confidence intervals (solid vertical lines) for OTUs $j = 16, 34$, and 49. Inference on the covariate effects is different from that under the proposed. For example, domoic acid (DA) levels do not have significant effects on the mean counts for OTU $j = 16$ and 49 from **Figures 7A,C**. Comparing **Figures 7A,C** to **Figure 7**, the NBMM produces wider interval estimates for $\beta_{j,p}$. Histograms of the MLEs of $\beta_{j,p}$, $\hat{\beta}^{\mathrm{NBMM}}_{j,p}$ under the NBMM are shown in Supplementary Figure 14. The histograms are much dispersed than those under the proposed model shown in Supplementary Figure 13. Estimates $\hat{\beta}_{j,p}$ and $\hat{\beta}^{\mathrm{NBMM}}_{j,p}$ for all covariates are also compared in Supplementary Figure 15. From the figure, the NBMM yields extremely large or small values for $\hat{\beta}_{j,p}$ for some OTUs, possibly due to the convergence problem. The insert plots show that for regions of small values of $\hat{\beta}_{j,p}$, the estimates under the proposed are more shrunken toward zero than those under the NBMM, similar to the results in section 3.1. The overdispersion parameter estimates under the NBMM tend to be smaller than those under the proposed (shown in Supplementary Figure 12f), which may lead to different predictive distributions of OTU counts.

4. DISCUSSION AND CONCLUSIONS

In this paper, we developed a Bayesian semiparametric regression model for joint analysis of microbiome data. We formulated the mean counts of OTUs as a product of factors and built models for the factors. We utilized the regularizing priors with mean constraints to avoid possible idenfiability issues, and the process convolution model to capture the temporal dependence structure in the baseline mean abundance of an OTU. The flexible model developed for baseline abundance enables joint analysis of all OTUs in the data and allows borrowing information across OTUs, across samples, and across time points. The model produces accurate estimates of the baseline mean counts and yields improved estimates of the effects of the covariates. We incorporated the Laplace distribution, a sparsity inducing shrinkage prior for the coefficients and the proposed model produces sparse estimates that is more desirable when the problem is high-dimensional and covariates are highly



$\hat{\beta}_{j,p}$ with their 95% credible intervals for OTU 16 ($j = 16$)

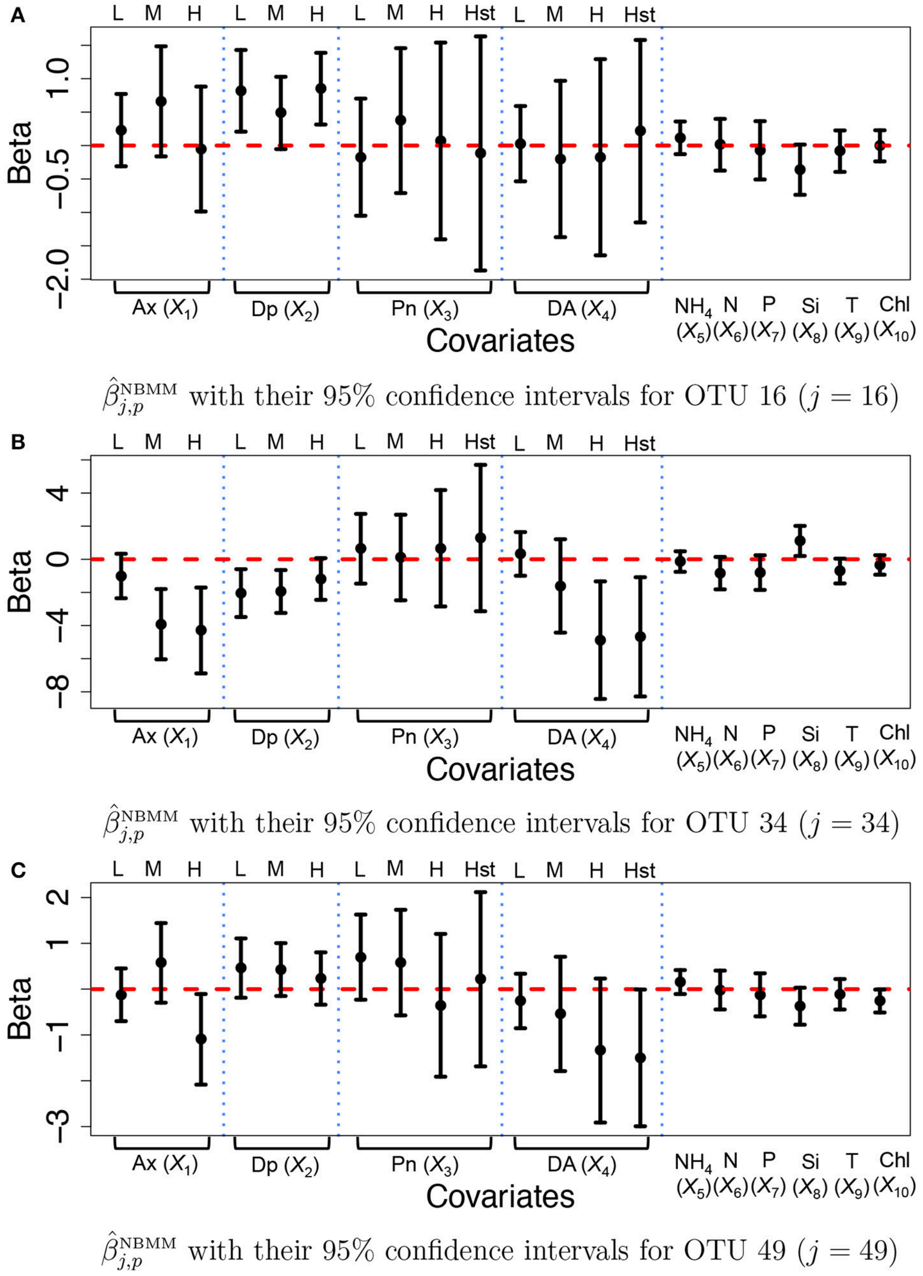
$\hat{\beta}_{j,p}$ with their 95% credible intervals for OTU 34 ($j = 34$)

$\hat{\beta}_{j,p}$ with their 95% credible intervals for OTU 49 ($j = 49$)

FIGURE 6 | Ocean microbiome data—proposed model. Posterior Inference on β_j for some selected OTUs ($j = 16, 34, 49$). Dots represent the posterior means $\hat{\beta}_{j,p}$ of $\beta_{j,p}$. Each vertical line connects the lower bounds and the upper bounds of 95% credible intervals.

correlated. We compared the proposed model to a comparable frequentist model that does separate analyses for individual OTUs. The comparisons through the simulation study and real data analysis show better performance of the proposed model.

Although we focused on the analysis of NGS count data, the proposed model is quite general and can be applied for analyses of any count data. Future work will explore alternative approaches to model the effects of covariates on the mean counts. For example, one may consider a nonparametric model using linear combinations of basis functions (Kohn et al., 2001) to flexibly capture shape in the response function. In such a case, an elaborate development of the prior model may be needed to produce a robust inference since both the baseline mean counts and the covariate effects are nonparametrically modeled. Other possible extensions are to include a variable selection method such as a stochastic search variable selection (George and McCulloch, 1993) if it is reasonable to assume that some covariate effects are exactly zero, and to let coefficients vary over time if covariate effects evolve with time. For time varying coefficients, we may use the random walk process in Leybourne (1993) to



$\hat{\beta}^{\text{NBMM}}_{j,p}$ with their 95% confidence intervals for OTU 16 ($j = 16$)

$\hat{\beta}^{\text{NBMM}}_{j,p}$ with their 95% confidence intervals for OTU 34 ($j = 34$)

$\hat{\beta}^{\text{NBMM}}_{j,p}$ with their 95% confidence intervals for OTU 49 ($j = 49$)

FIGURE 7 | Ocean microbiome data—NBMM. Inference on β_j for some selected OTUs ($j = 16, 34, 49$) under the negative binomial mixed model (NBMM). Dots represent the maximum likelihood estimates $\hat{\beta}^{\text{NBMM}}_{j,p}$ of $\beta_{j,p}$. Each vertical line connects the lower bounds and the upper bounds of 95% confidence intervals.

induce relationship between $\beta_{j,p,t-1}$ and $\beta_{j,p,t}$. Considering the high dimensionality in OTU data, posterior computation may need to be carefully handled. Also, prior information may be needed to produce sensible inference due to sparsity in data.

AUTHOR CONTRIBUTIONS

JL developed the statistical model and conducted simulation studies and data analysis. She also prepared the first draft and led the collaboration with MS-M for statistical analysis. MS-M provided the ocean microbiome data, participated the statistical model development, provided biological interpretation of the resulting inference and edited the manuscript.

FUNDING

This work was supported by NSF grant DMS-1662427 (JL) and NOAA-ECOHAB PROGRAM (Grant No. NA17NOS4780183, ECOHAB #905) (MS-M and JL). Collection of environmental data from the Santa Cruz Municipal Wharf was supported by Cal-PReEMPT with funding from the NOAA-MERHAB (#206), ECOHAB and IOOS programs.

ACKNOWLEDGMENTS

We gratefully acknowledge Raphael Kudela who provided the environmental data in this study, Michael Kempnich and Sanjin Mehic for doing the water sampling and processing the water samples for DNA extraction.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00522/full#supplementary-material

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

Banerjee, S. Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data. 2nd Edn.* Boca Raton, FL: CRC Press; Chapman & Hall.

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

Cressie, N. (1992). Statistics for spatial data. *Terra Nova* 4, 613–617.

George, E. I., and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *J. Am. Stat. Assoc.* 88, 881–889.

Gibbons, S. M., Kearney, S. M., Smillie, C. S., and Alm, E. J. (2017). Two dynamic regimes in the human gut microbiome. *PLoS Comput. Biol.* 13:e1005364. doi: 10.1371/journal.pcbi.1005364

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environ. Ecol. Stat.* 5, 173–190.

Higdon, D. (2002). "Space and space-time modeling using process convolutions," in *Quantitative Methods for Current Environmental Issues*, eds C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi (London: Springer), 37–56.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput.* 11, 313–322. doi: 10.1023/A:1011916902934

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Can. J. Stat.* 15, 209–225.

Leybourne, S. J. (1993). Estimation and testing of time-varying coefficient regression models in the presence of linear restrictions. *J. Forecast.* 12, 49–62.

Lee, H. K., Higdon, D. M., Calder, C. A., and Holloman, C. H. (2005). Efficient models for correlated data via convolutions of intrinsic processes. *Stat. Model.* 5, 53–74. doi: 10.1191/1471082X05st085oa

Li, Q., Guindani, M., Reich, B., Bondell, H., and Vannucci, M. (2017). A bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Mining* 10, 393–409. doi: 10.1002/sam.11350

Liang, W. W., and Lee, H. K. (2014). Sequential process convolution gaussian process models via particle learning. *Stat. Interface* 7, 465–475. doi: 10.4310/SII.2014.v7.n4.a4

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models, No. 37 in Monograph on Statistics and Applied Probability.* Boca Raton, FL: Chapman & Hall/CRC.

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

Needham, D. M., and Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat. Microbiol.* 1:16005. doi: 10.1038/nmicrobiol.2016.5

Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337

Polson, N. G., and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *J. R. Stat. Soc. Ser. B* 74, 287–311. doi: 10.1111/j.1467-9868.2011.01015.x

Robinson, M. D., and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887. doi: 10.1093/bioinformatics/btm453

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* 2:4. doi: 10.1186/2049-2618-2-4

Sison-Mangus, M. P., Jiang, S., Kudela, R. M., and Mehic, S. (2016). Phytoplankton-associated bacterial community composition and succession during toxic diatom bloom and non-bloom events. *Front. Microbiol.* 7:1433. doi: 10.3389/fmicb.2016.01433

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

Witten, D. M. (2011). Classification and clustering of sequencing data using a poisson model. *Ann. Appl. Stat.* 5, 2493–2518. doi: 10.1214/11-AOAS493

Woo, P., Lau, S., Teng, J., Tse, H., and Yuen, K.-Y. (2008). Then and now: use of 16s rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin. Microbiol. Infect.* 14, 908–934. doi: 10.1111/j.1469-0691.2008.02070.x

Xiao, S. (2015). *Bayesian Nonparametric Modeling for Some Classes of Temporal Point Processes*. Ph.D. thesis, University of California, Santa Cruz.

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

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 Microbiology

TECHNOLOGY REPORT
published: 23 April 2018
doi: 10.3389/fmicb.2018.00785

SplinectomeR Enables Group Comparisons in Longitudinal Microbiome Studies

Robin R. Shields-Cutler[1], Gabe A. Al-Ghalith[2], Moran Yassour[3,4] and Dan Knights[1,5]*

[1] BioTechnology Institute, College of Biological Sciences, University of Minnesota, Minneapolis, MN, United States, [2] Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, United States, [3] Broad Institute of Massachusetts Institute of Technology, Harvard University, Cambridge, MA, United States, [4] Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, [5] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Ekaterina Smirnova,
University of Montana, United States
Richard Allen White III,
RAW Molecular Systems (RMS) LLC, United States

****Correspondence:***
Dan Knights
dknights@umn.edu

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

Received: *15 November 2017*
Accepted: *06 April 2018*
Published: *23 April 2018*

Citation:
Shields-Cutler RR, Al-Ghalith GA, Yassour M and Knights D (2018) SplinectomeR Enables Group Comparisons in Longitudinal Microbiome Studies. Front. Microbiol. 9:785. doi: 10.3389/fmicb.2018.00785

Longitudinal, prospective studies often rely on multi-omics approaches, wherein various specimens are analyzed for genomic, metabolomic, and/or transcriptomic profiles. In practice, longitudinal studies in humans and other animals routinely suffer from subject dropout, irregular sampling, and biological variation that may not be normally distributed. As a result, testing hypotheses about observations over time can be statistically challenging without performing transformations and dramatic simplifications to the dataset, causing a loss of longitudinal power in the process. Here, we introduce splinectomeR, an R package that uses smoothing splines to summarize data for straightforward hypothesis testing in longitudinal studies. The package is open-source, and can be used interactively within R or run from the command line as a standalone tool. We present a novel in-depth analysis of a published large-scale microbiome study as an example of its utility in straightforward testing of key hypotheses. We expect that splinectomeR will be a useful tool for hypothesis testing in longitudinal microbiome studies.

Keywords: bioinformatics, microbiome analysis, R packages, computational biology methods, permutation tests, longitudinal data analysis

INTRODUCTION

Biological studies in humans are subject to significant variability and noise, often great enough to obscure all but the most dramatic differences. Longitudinal studies are powerful in these cases, allowing researchers to observe (and account for) both inter- and intra-individual variability, or measure changes in response to an intervention in real time (Gonzalez et al., 2012; Gerber, 2014). As the costs of DNA sequencing have decreased, microbiome researchers have a greater opportunity to perform such longitudinal studies. While longitudinal data with multiple timepoints always provide more information than single-timepoint data, the computational tools to analyze longitudinal microbiome studies with multiple timepoints per subject lag behind. A number of practical concerns often complicate analysis of longitudinal microbiome data: time points are usually not in sync or differ in number between subjects, longitudinal variation may not follow a normal distribution, and timeseries data may follow arbitrary curves, for example during the maturation of the infant microbiome. To overcome these challenges, researchers in many studies have collapsed samples across time points to average individuals' signals or they have summarized first with multivariate approaches that condense the initial observations (e.g., David et al., 2014; Zhou et al., 2015; Yassour et al., 2016). These approaches have been sufficient to make important discoveries in published studies, but there may still be opportunities to gain statistical power by using additional information content and directionality of the temporal axis.

To address this gap in data analysis, we introduce splinectomeR for direct hypothesis testing of categorical variables in longitudinal studies. SplinectomeR's implementation is straightforward and complements recently developed mixed-effects models that are used for discovering differentiating taxa (Chen and Li, 2016). At the core of the tests is the *loess* spline that uses weighted local polynomials to model data that may not follow any classical model or shape (as is common in real biological data) (Cleveland, 1979; Cleveland and Devlin, 1988). Null distributions are generated by permutation of the data, similar to methods implemented in multivariate tests such as PERMANOVA (Anderson, 2001). Lastly, its implementation as an R package makes it practical and easy to adopt by microbiome researchers.

splinectomeR contains three key functions: *permuspliner*, which tests for an overall significant difference between two groups' summary splines over the longitudinal time course; *sliding_spliner*, which interpolates across the group spline and tests for significance between two groups at each interval to illuminate regions of time where significant differences exist; and *trendyspliner*, which tests for a significant non-zero overall trend in a single population over time. To demonstrate the utility of these tests, we performed additional testing of key hypotheses in a large published cohort of 37 infant microbiomes sampled over the first 3 years of life (Yassour et al., 2016). SplinectomeR is open-source and freely available for download and installation on GitHub at https://github.com/RRShieldsCutler/splinectomeR.

METHODS AND IMPLEMENTATION

splinectomeR contains three primary functions that test specific hypotheses about longitudinal trends (see **Figure 1** for schematic diagram). Each function uses *loess* splines to smooth longitudinal data before performing the specific statistical test. The input is a properly formatted data frame: each quantitative measurement or metadata category is its own column, and each row is a separate observation. This is a common structure for bioinformatics metadata, including in microbiome analyses, and therefore tests may be performed with little or no reformatting required. The standalone command-line version of the scripts requires a tab-delimited file of the same structure. We refer the reader to the online package vignettes (included as Supplemental Data Files 1 and 2) for examples of proper input data and for manual data wrangling techniques.

Permuspliner

The objective of the permuspliner function is to test whether two groups of individuals follow more different trajectories over time (or along any continuous axis) than would be expected by random chance. It compares two groups over time without collapsing the timepoints to a single averaged point. Since differences between groups may not be consistent over time, and responses may even invert relative to one another over time, collapsing could mask or nullify important differences between groups. This problem can be avoided by considering the entire time series in a statistical comparison. Permuspliner fits a spline to the average time series of each group of individuals, and then measures the absolute area between the splines to determine the observed group difference.

Specifically, the input data is first split into the two groups to be compared, A and B. In some studies, participants may only appear once or twice in the dataset; to filter out samples from low-prevalence participants, the user may set a threshold with "*cut_low = n*." A *loess* spline is fit to each group's total time series, if it also meets the minimum data sparseness threshold set by the "*cut_sparse = n*" parameter. Lastly, the direction of the null hypothesis can be set with "*test_direction = [more/less]*"; this enables the user to test for not only group differences that are greater than expected by chance (*test_direction = "more,"* the default) but also differences smaller than expected by chance (*test_direction = "less"*). To calculate the observed group distance, points are interpolated along each spline and the sum of the areas across these points yields the group distance. Splinectomer uses 1000 points by default but this can be edited in the function arguments (with "*ints = n*") to account for a longer time series or greater/lower resolution. This area determination is then recalculated after permuting the group assignments: to which group each participant belongs (A vs. B) is shuffled randomly without replacement. Permuting the group labels, as opposed to the participant data points, preserves the underlying distributions and patterns of the individuals' timeseries curves. The permutation is repeated (default setting is 999 permutations) to generate a null distribution over the random between-group distances, from which an empirical *p*-value is calculated and reported by comparison to the observed distance. Because the random distribution is built from the observed values, it inherently reflects the noise and variability of the dataset, and is therefore tailored to each study's unique character. SplinectomeR also includes plotting functions for visualizing the permuted distribution.

Sliding Spliner

As a complement to the permuspliner test, the sliding spliner function allows the user to test the data series at defined intervals and ask whether the two groups of interest are significantly different at a any point in time during the time series. This is often challenging in clinical datasets where sampling is not coordinated between individuals, so analysis requires artificial binning or averaging across time in order to compare enough data at any one time point. Here, each individual—as opposed to the whole group, as in the permuspliner test above—is summarized by a spline, thus filling gaps across their time series. Every interpolated interval can then be tested as a complete distribution of all participants without binning or unnecessarily dropping samples. Low-prevalence participants are again filtered out with the "*cut_low*" parameter. Since the splines are not extrapolated beyond the start and end of the individual's time

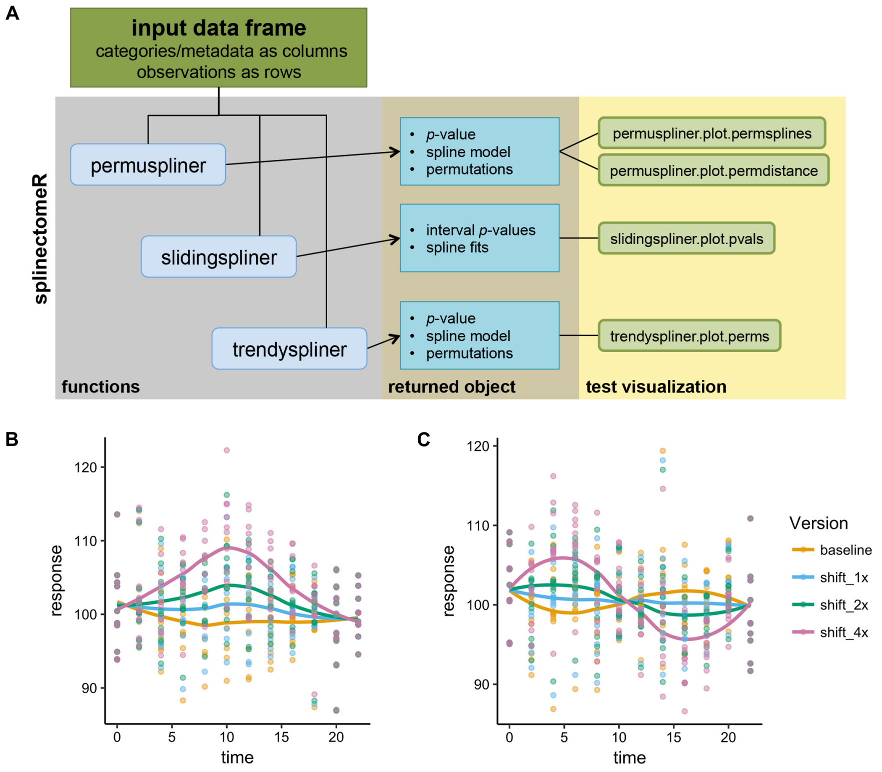


FIGURE 1 | Schematic of the splinectomeR analysis package and simulated analyses. **(A)** Each of the three primary tests produces a results list object containing several data articles including the *p*-value(s), and which can also be used to generate pre-configured data visualizations. To demonstrate that the splinectomeR package detects non-linear changes between groups during a time series, data were simulated for 10 individuals over time and perturbed at three magnitudes (1x, 2x, 4x). The perturbation was done at one **(B)** or two **(C)** regions of the time series The permuspliner test finds that these changes are less likely to be random as the magnitude increases, as would be expected. In **(B)**, *p* = 0.3, *p* = 0.005, and *p* = 0.001 for 1x, 2x, and 4x shifts, respectively; in **(C)** *p* = 0.79, *p* = 0.36, and *p* = 0.002 for 1x, 2x, and 4x shifts, respectively.

series, the start and end regions may become less dense as fewer subjects have early and/or late samples. To account for this, the "*test_density* = *n*" argument sets the minimum number of participants required in each group to perform and report a *p*-value (default = 3). The output from this function is a table containing a *p*-value for each time interval, summarized by the companion plotting function where intervals at which the groups significantly diverge are visualized, and the points are scaled according to the data density—larger circles on the plot mean more data were used to calculate the *p*-value at that interval than for smaller circles.

Trendyspliner

While testing for association between a categorical variable and a longitudinal variable is straightforward using regressions and correlations, methods for quickly testing whether a response is increasing or decreasing over time are less established. The trendyspliner function tests whether a set of responses in one group changes in a non-zero direction over the time series (or other continuous independent axis). A spline is fit to the data and interpolated across the number of set intervals. Non-zero change is measured as the area between the group spline and a linear baseline that is established from the start point of the group spline. Thus, if the pattern of observations does not meaningfully increase or decrease over time, the spline will not diverge from the baseline and the areas will remain small. To generate the null distribution, the time series within each individual is then permuted, and the spline is recalculated along with the area to the permuted baseline. The permutation is repeatedly executed to generate a random distribution of areas from which the two-sided *p*-value is calculated by comparison to the observed value. In some biological measurements, individuals' initial values may be variable (e.g., body weight, height). To normalize these differences and improve the ability to detect an increase or decrease in these situations, the user can elect to "*mean_center*" the observations before calculations, which shifts each individuals' mean over time to the group mean from all individuals. As in the other modes, a plotting function allows the user to visualize the permuted splines in the context of the real data.

User Customization

In each of the above functions, the user can alter several additional parameters for specific applications. These include the spline span parameter, a standard spline parameter that determines how large the local smoothing neighborhood is, and therefore the degree of smoothing; the number of permutations, which influences the sensitivity of the test and *p*-value, as more permutations will allow a lower *p*-value to be detected but will also increase run time (default is 999 permutations); and the number of intervals over which to divide the data (larger values also increase run time and memory but provide finer resolution and more precise comparisons). These arguments allow flexibility for more advanced users with particular needs and unique data shapes.

Package Implementation

The splinectomeR code was written in R version 3.4.0, and the package was built in RStudio using devtools and roxygen2 to generate and populate the package documentation (Wickham and Chang, 2017; Wickham et al., 2017). SplinectomeR is open-source and freely available on GitHub at https://github.com/RRShieldsCutler/splinectomeR. The figures generated by the secondary plotting functions are ggplot2 objects and can be saved in most image formats at a size and resolution specified by the user (e.g., through base R drivers or the "ggsave" function) (Wickham, 2009).

EXAMPLE ANALYSIS

To demonstrate how splinectomeR detects group changes over a time series, we first generated a simulated data set. The response variable was perturbed at one or two regions of the time series, and at three magnitudes of change (**Figures 1B,C**). A linear model is a poor fit to these data shapes, and testing for absolute change from beginning to end is not sensitive to the internal dynamics. However, splinectomeR uses the entire dataset to compare between the baseline and perturbed data. As the magnitude of change increases, the permuspliner test yields decreasing *p*-values as expected in both the single region ($p = 0.3$, $p = 0.005$, $p = 0.001$ for 1x, 2x, and 4x shifts, respectively; **Figure 1B**) and double region perturbations ($p = 0.79$, $p = 0.36$, $p = 0.002$ for 1x, 2x, and 4x shifts, respectively; **Figure 1C**). Non-linear changes of this sort during a time series may be of great biological interest, although their statistical significance is difficult to test with existing tools.

Freely available datasets were used to test and further demonstrate the splinectomeR functions, as documented in the package vignettes that are available to view online in HTML format at https://rrshieldscutler.github.io/splinectomeR/. A proof-of-concept analysis was performed on the ChickWeights dataset in the R "datasets" package, and is available as a vignette on the website. To demonstrate splinectomeR's utility on a more complex dataset, we used the publicly available OTU tables and associated metadata from a published longitudinal study of infant microbiomes by Yassour et al. (2016).

Analysis of Longitudinal Microbiome Data

We tested splinectomeR's utility on the taxonomic and metadata profiles from Yassour et al. (2016), to evaluate whether we could statistically support patterns described by the authors and investigate novel hypotheses. A more extensive analysis including all code used for data formatting is available as an online vignette and as Supplemental Data File 2.

In several figures in the original publication, the authors draw visual comparisons between taxon abundances in antibiotic exposed versus non-exposed infants. These stream plots are powerful in displaying the inter-individual diversity and inspired us to use splinectomeR to perform statistical hypothesis testing on the time series data for differences between the infant groups. SplinectomeR's permuspliner function can test whether a taxon's abundance pattern over time is significantly different between antibiotic-exposed and non-exposed infants. We used splinectomeR to calculate that the difference in *Bacteroidaceae* abundance is not statistically significant across the overall time series ($p = 0.28$). The permutated differences support this conclusion, as the output shows in **Figure 2A**. However, the results indicate that the groups may be diverging toward the end of the time series, as the observed distance is higher relative to most of the permuted values. We tested this with the sliding spliner function, which generates a series of *p*-values across the longitudinal scale. The results, as the function's output plot shows in **Figure 2B**, indeed show that there is a temporal pocket of significance surrounding the 30-month time point. We were also able to confirm the finding that the genus *Bacteroides* is significantly different between vaginal and cesarean born infants ($p = 0.04$), and that this difference is most pronounced in the first year of life (see Supplemental Data File 2).

Because these tests are implemented as R functions, they can be used programmatically for multiple hypothesis testing, such as testing all dominant bacterial families for significant differences between antibiotic exposure status. We performed this test on the present dataset (see vignette for full analysis), revealing that *Porphyromonadaceae* is the most discriminatory family ($p = 0.05$). The built-in plotting function (permuspliner.plot.permsplines) shows that antibiotic-exposed infants do indeed develop a notably higher abundance of *Porphyromonadaceae* (**Figure 2C**). The permuted splines lie predominantly between the two observed curves, supporting the conclusion that this difference is larger than expected by chance, and this observation becomes greater over time. The distance plot further supports this conclusion (**Figure 2D**), showing that initial distances are not greater than chance but become significant after approximately 12 months of age.

In their published analysis, the authors found a significant difference in abundance of the butyrate-producing *Clostridium* groups IV and XIV by antibiotic status at the final time point (36 months). Given butyrate's important roles in gut homeostasis (e.g., Pryde et al., 2002; Ridaura et al., 2013; Zhang et al., 2016), it is worth investigating whether this difference exists over the infants' first 3 years of development, or is just established at 36 months. To test this, we used the permuspliner function on a summarized table containing the following putatively butyrate-producing genera from *Clostridium* groups IV and XIV present in the OTU table: *Clostridium*, *Coprococcus*, *Dorea*, *Lachnospira*, *Roseburia*, *Ruminococcus*, and *Faecalibacterium* (Louis and Flint, 2009; Lopetuso et al., 2013; Van den Abbeele et al., 2013). When all of the longitudinal

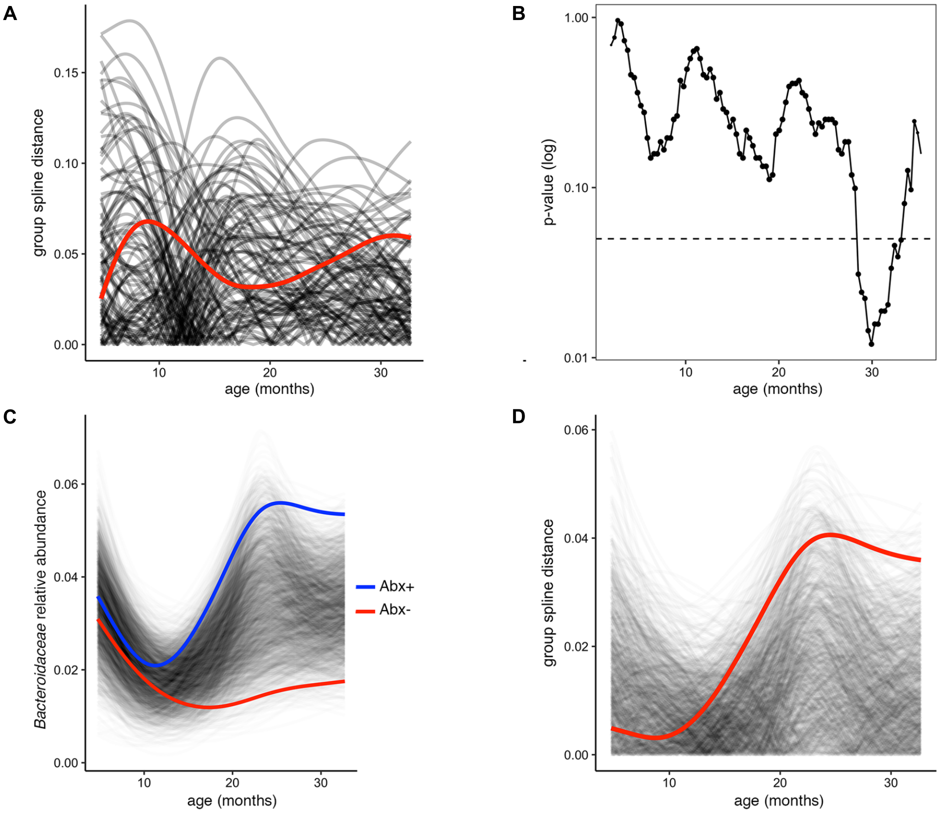


FIGURE 2 | Permuted spline tests for statistical significance in longitudinal microbiome data. **(A)** The permuspliner output plot shows that the difference between *Bacteroidaceae* abundance between antibiotic exposed and non-exposed infants (distance between group splines shown as red line) is not significantly greater than 95% of the permuted values shown in translucent black. **(B)** The plot output of the sliding spliner function shows the *p*-value at each specified interval (shown with default 100 intervals) derived from the distribution of points from individuals' smoothed splines. Dotted line indicates $p = 0.05$. The number of infants with data at a given interval is used to scale the data point size, as some entered and exited the study later or earlier, respectively. **(C)** *Porphyromonadaceae* abundance over time distinguishes antibiotic exposed (group spline in blue) and non-exposed infants (group spline in red; 999 permutations, $p = 0.053$). **(D)** Group distance plot, as in **(A)**, for the *Porphyromonadaceae* comparison, showing that permuted splines support a trend toward a greater true statistical difference after 12 months of life.

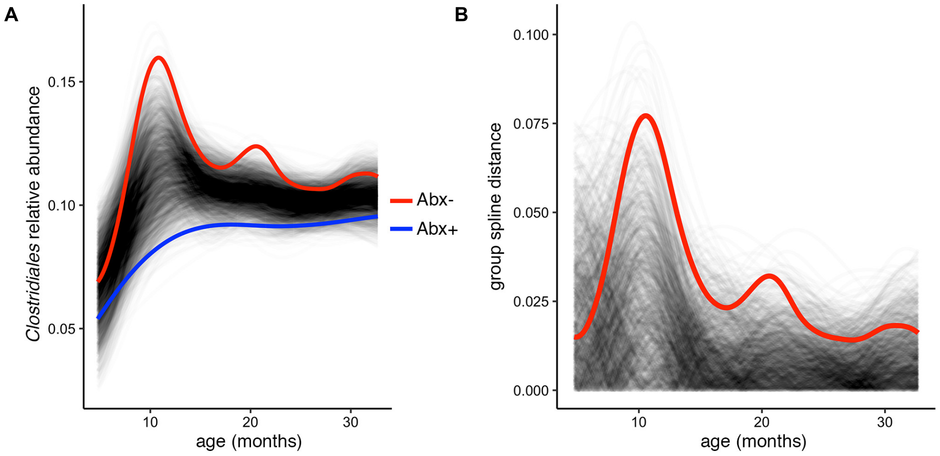


FIGURE 3 | Complete time series analysis highlights a significant and temporally maintained deficiency in butyrate-producing *Clostridiales* among infants exposed to antibiotics. **(A)** Results plot generated by the permuspliner test, showing enriched abundance of *Clostridium* groups IV and XIV in infants not exposed to antibiotics (Abx–, red line). **(B)** Corresponding distance plot output, showing that the observed difference between the groups exceeds the random permuted distribution 997/1000 times, which supports the statistically significant finding ($p = 0.003$).

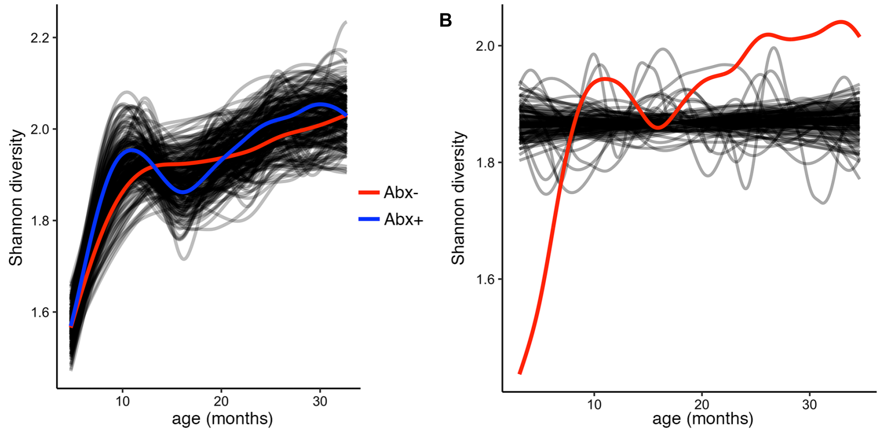


FIGURE 4 | Alpha diversity increases over time but is not different between antibiotic exposure groups. (**A**) Output plot from the permuspliner test showing that Shannon diversity is not significantly different in infants exposed to antibiotics (Abx+, blue) compared to those who were not (Abx–, red), $p = 0.96$. **(B)** The results plot from the trendyspliner function shows that the permuted data form a zero-change distribution from which the real data (red line) is significantly distinct ($p = 0.001$). This supports the hypothesis that alpha diversity increases over time in the first 3 years of the infants' lives.

data are included in this comparison, we find that antibiotic exposed infants do indeed have significantly lower *Clostridium* group IV/XIV abundance compared to non-exposed infants over time (**Figure 3A**). Notably, our test using the entire 36-month time series yields a lower *p*-value than that reported using just the 36 month data: $p = 0.003$ vs. $p = 0.037$, respectively, and the resulting plot suggests that the most divergent time point is near 1 year of age (**Figures 3A,B**). This demonstrates the utility and simplicity of the splinectomeR tests; we are able to directly include 3 years of observations, strengthen support for this finding, and suggest directions for new hypotheses.

To test and demonstrate the third function in the splinectomeR package, we hypothesized that the infants' alpha diversity would significantly increase over time. We can use the permuspliner test to show that the alpha diversity is highly similar between the antibiotic ($p = 0.92$) and birth mode ($p = 0.98$) groups, but to statistically test whether it changes over time, we used the trendyspliner function. As shown in **Figure 4**, the randomly permuted splines generate a linear and zero-slope distribution, while the observed spline increases steadily over the time series, confirming our hypothesis that infant microbiome alpha diversity increases over time ($p = 0.01$). In this case, the trend is evident and expected; many biological datasets, however, have slight trends that are hard to interpret, in which case the trendyspliner test provides a straightforward permutation-based statistical test to confirm whether the deviation is greater than expected from chance.

Summary

Yassour et al. (2016) present a prodigious dataset with dense longitudinal data that details the human gut microbiome's complex dynamics over the first 3 years of life. The approaches presented above provide statistical support for observations and conclusions the authors reported, and allow us to test and develop additional hypotheses from the dataset. Researchers analyzing new longitudinal microbiome data with multiple samples per subject may benefit from these analytical capabilities provided by splinectomeR.

DISCUSSION

Longitudinal studies hold great promise for understanding the effects of interventions and environmental stimuli in the context of a naturally variable population. Analysis of these complex data has been impeded by a lack of clear, simple methods for directly comparing observations across multiple individuals and many time points without averaging or summarizing across time points. As we have demonstrated here using a large-scale longitudinal microbiome study, the splinectomeR package performs straightforward tests that are easy to interpret and will allow researchers to test important hypotheses from within R or a command line interface.

The approaches here are not without limitations; reliance on the *loess* spline means that the tests may be impacted by outliers, particularly in sparse datasets. User-definable arguments for sparseness cut offs and spline resolution (the smoothing parameter or spar) can help minimize these effects. Additionally, the size of longitudinal studies mean that the tests with many permutations can be slow to complete, though still easily performed on a standard workstation running R. From a standard metadata table, splinectomeR tests are run with a single command in R or on the command line. We provide the user with interpretable results in the form of pre-formatted plots that can be saved at publication quality and re-generated from the results object stored by the function.

CONCLUSION

In summary, we have shown how a new open-source R package, splinectomeR, can quickly assess statistical significance in large longitudinal microbiome studies by summarizing longitudinal group data with splines and using randomly permuted distributions to evaluate the probability that the observed magnitude of differences between groups, or of trends over time, is due to chance. By providing three distinct types of hypothesis tests, we can explore overall changes in abundances or other biological measures, and compare longitudinal trends between groups of interest. Altogether, we are able to perform these tests in a way that takes full advantage of longitudinal data and maintains individual observations, thus leveraging all possible data points. Longitudinal studies generating "big data" with multi-omics approaches are now commonplace, but we lack appropriate tools to interpret these data. We offer splinectomeR as an open-source solution to testing key hypotheses in complex longitudinal data. SplinectomeR may also simplify analysis for longitudinal studies beyond the microbiome research field.

AUTHOR CONTRIBUTIONS

RS-C, GA-G, and DK conceived and designed the spline tests. RS-C and GA-G wrote the R code, and RS-C built the R package and conceived, and performed the example experiments. RS-C and MY analyzed the experiments. RS-C and DK wrote the manuscript.

FUNDING

This research was supported, in part, by NIH Grant R01AI121383 (to DK) and Minnesota Partnership for Biotechnology and Genomics Grant #13.28 (to DK).

ACKNOWLEDGMENTS

We are grateful to all members of the Knights lab for testing preliminary code and providing feedback on various aspects of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00785/full#supplementary-material

REFERENCES

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74, 829–836. doi: 10.1080/01621459.1979.10481038

Cleveland, W. S., and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* 83, 596–610. doi: 10.1080/01621459.1988.10478639

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

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

Gonzalez, A., King, A., Robeson, M. S., Song, S., Shade, A., Metcalf, J. L., et al. (2012). Characterizing microbial communities through space and time. *Curr. Opin. Biotechnol.* 23, 431–436. doi: 10.1016/j.copbio.2011.11.017

Lopetuso, L. R., Scaldaferri, F., Petito, V., and Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* 5:23. doi: 10.1186/1757-4749-5-23

Louis, P., and Flint, H. J. (2009). Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol. Lett.* 294, 1–8. doi: 10.1111/j.1574-6968.2009.01514.x

Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S., and Flint, H. J. (2002). The microbiology of butyrate formation in the human colon. *FEMS Microbiol. Lett.* 217, 133–139. doi: 10.1111/j.1574-6968.2002.tb11467.x

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., et al. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* 341:1241214. doi: 10.1126/science.1241214

Van den Abbeele, P., Belzer, C., Goossens, M., Kleerebezem, M., De Vos, W. M., Thas, O., et al. (2013). Butyrate-producing *Clostridium* cluster XIVa species specifically colonize mucins in an *in vitro* gut model. *ISME J.* 7, 949–961. doi: 10.1038/ismej.2012.158

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wickham, H., and Chang, W. (2017). *devtools: Tools to Make Developing R Packages Easier*. Available at: https://CRAN.R-project.org/package=devtools

Wickham, H., Danenberg, P., and Eugster, M. (2017). *roxygen2: In-Line Documentation for R*. Available at: https://CRAN.R-project.org/package=roxygen2

Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.-M., Härkönen, T., Ryhänen, S. J., et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* 8:343ra81. doi: 10.1126/scitranslmed.aad0917

Zhang, Q., Wu, Y., Wang, J., Wu, G., Long, W., Xue, Z., et al. (2016). Accelerated dysbiosis of gut microbiota during aggravation of DSS-induced colitis by a butyrate-producing bacterium. *Sci. Rep.* 6:27572. doi: 10.1038/srep27572

Zhou, Y., Shan, G., Sodergren, E., Weinstock, G., Walker, W. A., and Gregory, K. E. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLoS One* 10:e0118632. doi: 10.1371/journal.pone.0118632

Conflict of Interest Statement: DK serves as CEO and holds equity in CoreBiome, a company involved in the commercialization of microbiome analysis. The University of Minnesota also has financial interests in CoreBiome under the terms of a license agreement with CoreBiome. These interests have been reviewed and managed by the University of Minnesota in accordance with its Conflict-of-Interest policies.

The other 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 Microbiology

ORIGINAL RESEARCH
published: 22 May 2018
doi: 10.3389/fmicb.2018.01037

On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities

Brandie D. Wagner[1,2*], *Gary K. Grunwald*[1], *Gary O. Zerbe*[1], *Susan K. Mikulich-Gilbertson*[3], *Charles E. Robertson*[4], *Edith T. Zemanick*[2] *and J. Kirk Harris*[2]

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, CO, United States, [2] Department of Pediatrics, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, United States, [3] Department of Psychiatry, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, United States, [4] Department of Molecular, Cellular and Developmental Biology, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, United States

Identification of the majority of organisms present in human-associated microbial communities is feasible with the advent of high throughput sequencing technology. As substantial variability in microbiota communities is seen across subjects, the use of longitudinal study designs is important to better understand variation of the microbiome within individual subjects. Complex study designs with longitudinal sample collection require analytic approaches to account for this additional source of variability. A common approach to assessing community changes is to evaluate the change in alpha diversity (the variety and abundance of organisms in a community) over time. However, there are several commonly used alpha diversity measures and the use of different measures can result in different estimates of magnitude of change and different inferences. It has recently been proposed that diversity profile curves are useful for clarifying these differences, and may provide a more complete picture of the community structure. However, it is unclear how to utilize these curves when interest is in evaluating changes in community structure over time. We propose the use of a bi-exponential function in a longitudinal model that accounts for repeated measures on each subject to compare diversity profiles over time. Furthermore, it is possible that no change in alpha diversity (single community/sample) may be observed despite the presence of a highly divergent community composition. Thus, it is also important to use a beta diversity measure (similarity between multiple communities/samples) that captures changes in community composition. Ecological methods developed to evaluate temporal turnover have currently only been applied to investigate changes of a single community over time. We illustrate the extension of this approach to multiple communities of interest (i.e., subjects) by modeling the beta diversity measure over time. With this approach, a rate of change in community composition is estimated. There is a need for the extension and development of analytic methods for longitudinal microbiota studies. In this paper, we discuss different approaches to model alpha and beta diversity indices in longitudinal microbiota studies and provide both a review of current approaches and a proposal for new methods.

Keywords: microbiome, Hill's numbers, repeated measures, alpha diversity, beta diversity, Shannon index, mixed model

OPEN ACCESS

Edited by:
*Jessica Galloway-Pena,
University of Texas MD Anderson
Cancer Center, United States*

Reviewed by:
*Bradley Stevenson,
University of Oklahoma, United States
Richard R. Rodrigues,
Oregon State University, United States*

***Correspondence:**
*Brandie D. Wagner
brandie.wagner@ucdenver.edu*

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

Received: *23 January 2018*
Accepted: *01 May 2018*
Published: *22 May 2018*

Citation:
*Wagner BD, Grunwald GK, Zerbe GO, Mikulich-Gilbertson SK, Robertson CE, Zemanick ET and Harris JK (2018) On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities.
Front. Microbiol. 9:1037.
doi: 10.3389/fmicb.2018.01037*

INTRODUCTION

Identification of the majority of organisms present in human-associated microbial communities is now feasible with the advent of high throughput sequencing technology. Several studies have shown large subject-to-subject variability (Flores et al., 2014) as well as many different factors that might contribute to variability in microbiome studies, i.e., diet, region, exposure, genetics, etc. Given the highly personalized microbiome, valuable information is likely to come from studies following subjects over time. The use of longitudinal study designs is important to better understand the contribution of the microbiome to human health (Flores et al., 2014). Complex study designs with longitudinal sample collection require analytic approaches to account for this additional source of variability, and to allow examination of changes within subjects.

Extension and development of analytic methods are needed for longitudinal microbiota studies (Gerber, 2014). Current approaches include extending models applied to individual taxa to address repeated measures over time (Chen and Li, 2016; Fang et al., 2016; Wagner et al., 2017) but not much attention has been given to discussion and extension of ecological community indices, which are useful for describing the community biodiversity. The majority of analytic methods for these measures were developed for studying one community over time in the field of ecology. This paper, therefore, focuses on the application and development of methods for diversity indices in order to model multiple communities (i.e., subjects) over time.

Several measures of diversity have been widely applied to microbiota data. The selection of a diversity measure is important as the inferences made can differ widely depending on the measure chosen (Jost, 2006; Ellison, 2010; Tuomisto, 2010a,b; Jurasinski and Koch, 2011; Moreno and Rodriguez, 2011; Tuomisto, 2011) and several analyses include multiple measures which makes consolidating the results challenging. For alpha diversity, the calculation and comparison of diversity curves (Renyi, 1961; Whittaker, 1972; Hill, 1973; Carranza et al., 2007; Studeny et al., 2011; Gotelli and Chao, 2013) has been proposed, which alleviates the need to choose a single diversity index. These curves provide a useful visualization but there currently is no method available to make inferences about the changing shape of the curves over time.

Furthermore, it is possible that no change in alpha diversity (single community/sample) may be observed despite the presence of a highly divergent community composition. Thus, it is also important to use a beta diversity measure (similarity between multiple communities/samples) that captures changes in community composition. Ecological methods developed to evaluate temporal turnover have currently only been applied to investigate changes of a single community over time (Collins et al., 2000; Korhonen et al., 2010; Yuan et al., 2016; Lewthwaite et al., 2017). In order to evaluate changes over time in multiple communities (i.e., subjects), an extension to a hierarchical model is needed.

In this paper, we discuss different approaches to model diversity indices in longitudinal microbiota studies. All approaches are illustrated using a motivating example described in section Description of Motivating Example. In section Single Alpha Diversity Index, a linear mixed model (also called a hierarchical model) is used to separately model three alpha diversity measures over time and the results are compared across measures. The recently proposed alpha diversity curves are explained in section Alpha Diversity Curves and we develop a hierarchical model approach to analyze these curves longitudinally with a non-linear mixed model. In section Beta Diversity, a description of how to model beta diversity in longitudinal studies is provided. This work provides both a review of current approaches and presents newly developed methods.

DESCRIPTION OF MOTIVATING EXAMPLE

The motivating example used throughout this paper is a prospective study of 50 subjects aged 10–22 years with cystic fibrosis (CF) and admitted for intravenous (IV) antibiotic therapy for a pulmonary exacerbation (Pex). All subjects were treated following standard clinical guidelines, at the discretion of their physician. Study evaluation and specimen collection occurred at three times, hospital admission (day 0–2; Beg Pex), hospital discharge (day 6–21; End Pex), and a clinical follow-up visit post-exacerbation (within 30 days of completing IV antibiotic treatment; Post Pex). A total of 123 sputum samples were collected and frozen prior to analysis: 31 subjects provided samples at all three times, 12 subjects missed 1 sample collection, and 7 subjects missed 2 sample collections. All models used for the analysis of this dataset assume data are missing at random. Written informed consent was obtained from all patients aged 18 years or older and from parents/legal guardians for patients under 18 years of age, and assent was obtained from patients aged 10–17 years. The study was approved by the Colorado Multiple Institutional Review Board (COMIRB #07-0365).

Bacterial profiles were determined by broad-range amplification and sequence analysis of 16S rDNA following previously described methods and validated in prior publications (Hara et al., 2012; Markle et al., 2013; Zemanick et al., 2017). Quality control procedures were performed on paired-end sequences (Zemanick et al., 2017). Assembled sequences were aligned and classified at the lowest taxonomic level with SINA version 1.2.11 (Pruesse et al., 2012) using the SILVA111 database (Quast et al., 2013) as reference configured to yield the SILVA taxonomy (www.arb-silva.de). Sorted paired-end sequence data were deposited in the National Center for Biotechnology Information Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under accession number SRP143768. Operational taxonomic units (OTUs) were produced by clustering sequences with identical taxonomic assignments (generally genus level groups). This process generated 20,183,481 sequences for 361 samples (average sequence length: 316 nt; average sample size: 83,722 sequences/sample; minimum sample size: 2,188; maximum sample size: 422,831). The median Goods coverage score was $\geq$ 99.25% at the rarefaction point of 2,188 (the minimum number of sequences for all samples). The software package Explicet version 2.10.5 (www.explicet.org) (Robertson et al., 2013) was used for calculation of diversity indices at the rarefaction point. Taxonomic data utilized in this analysis have been included as Supplementary Material and represent a subset of data from the parent study (excluding saliva samples and samples from repeated Pex events).

SINGLE ALPHA DIVERSITY INDEX

Diversity, defined as the description of "the variety and abundance of species in a defined unit of study," (Magurran, 2004) is a measure often used to describe the complexity of a community. Several measures of diversity have been widely applied to microbiota data and have been used previously as outcomes in longitudinal models (Gajer et al., 2012; Flores et al., 2014; Wagner et al., 2017). In this section we similarly apply linear mixed models to three diversity measures over time. These results serve as a useful comparator for the remaining sections of this paper.

Differences in Weights for Evenness and Richness Components Across Measures Explain Differences in Results

Diversity indices applied to microbiota data consist of differing weights of two components, richness and evenness (Jost, 2006). Richness is a count of the number of different taxa observed in the community without regard to their frequencies, and evenness refers to the equitability of the taxa frequencies in a community. Three commonly used alpha diversity measures include species observed, Shannon index and Simpson index:

$$\begin{aligned} S\left(obs\right) &= \sum\nolimits_k I\left(p_k > 0\right) \\ Shannon &= -\sum\nolimits_k p_k \, ln\left(p_k\right) \\ Simpson &= \sum\nolimits_k p_k^2, \end{aligned}$$

where p is some function of frequency, often relative abundance (proportion of total sequences) for each taxon, k.

Species observed is equal to richness and therefore provides no weight to the evenness component, Shannon index equally weights richness and evenness and Simpson index provides more weight to evenness (Jost, 2006). Moreover, the units are different across the measures, species observed is a count, Shannon index contains a logarithmic value and Simpson index is a sum of squared proportions. These differences in weighting and units explain differences often observed in results from each measure.

Motivating Example

Species observed, Shannon diversity index and Simpson diversity, as well as the corresponding evenness components, were separately modeled over time in CF patients during a Pex using a linear mixed model that included a random subject intercept with SAS PROC MIXED software. All three diversity measures show a decrease at the end of the Pex (hospital discharge), followed by an increase at follow-up, although measures still remained lower at follow-up than at the beginning of the Pex (**Figure 1**). Despite this agreement in general trends, the pairwise comparisons of times differ across the measures. The means at each time (**Table 1**) differed significantly across all three times for species observed and Shannon index ($p < 0.01$), but Simpson diversity differed only marginally across times ($p = 0.07$). Neither of the evenness measures change significantly over time.

Issues of Numerous Measures

Although the concept of diversity is rather straightforward, its application can be complicated for several reasons: (1) there are numerous commonly used diversity indices which can yield different results; (2) the nomenclature currently in use to describe diversity is complex and confusing; (3) partitioning diversity into components, such as richness and evenness, may be useful, but varies depending on the diversity measure; and (4) the application to sequence data is complicated by incomplete sampling, i.e., not all bacterial sequences may be measured due to differences in sequencing depth. These issues result in debates and general confusion over which diversity measure to use, misinterpretation of results, and an inability to compare results across studies.

Often these indices are incorrectly treated as interchangeable measures of the same characteristic without consideration of the variations in the mathematical properties of each diversity index. The measurement of diversity has been discussed in the ecological literature (Jost, 2006; Ellison, 2010; Jurasinski and Koch, 2011; Moreno and Rodriguez, 2011; Tuomisto, 2011) and there has been an acknowledgement within the field that more rigor is needed. One approach is the calculation and comparison of diversity curves (Renyi, 1961; Whittaker, 1972; Hill, 1973; Carranza et al., 2007; Studeny et al., 2011; Gotelli and Chao, 2013) which provides information across multiple weights of the components of richness and evenness and alleviates the need to choose a single diversity index.

ALPHA DIVERSITY CURVES

The computational formula for diversity curves is

$$D_{(q)} = \left(\sum\nolimits_{k=1}^{K} {p_k}^q\right)^{\frac{1}{1-q}},$$

where D is most commonly calculated for $q = 0, 1, 2$ and p is some function of frequency, often relative abundance (proportion of total sequences) for each taxon, k, when applied to sequencing data. D is undefined for $q = 1$, so the limit as q approaches 1 is used instead.

In this equation, the order, q, determines how much weight is given to abundant vs. rare taxa (evenness). Species observed ($q = 0$) weights rare taxa more heavily since the abundance of each taxon is not considered. Conversely with diversity of orders > 1 (e.g., Simpson $q = 2$), more weight is given to the more abundant species. Only when $q = 1$ [Shannon index, specifically exp(Shannon index)] are the rare and abundant species equally weighted (Jost, 2006).

A plot of D vs. varying values of q can provide a more complete way to convey diversity of a community compared

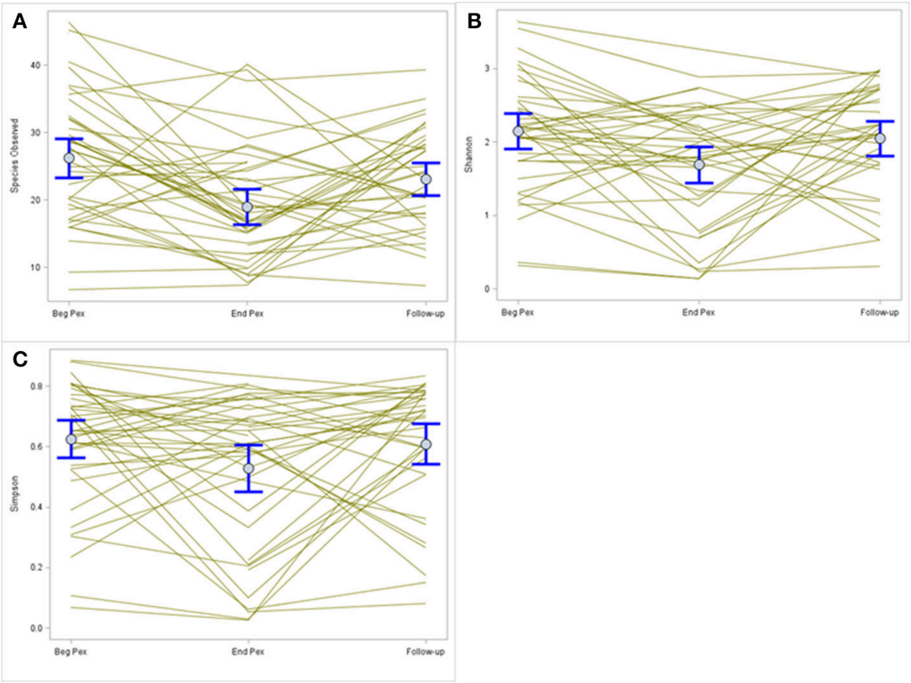


FIGURE 1 | Comparison of alpha diversity over time. Species observed **(A)** shows a decrease in values after completion of IV antibiotic treatments that increase at follow-up. A similar pattern is observed for the Shannon and Simpson diversity indices **(B,C,** respectively) but the magnitude of change differs for each index.

to using a single measure (Tothmeresz, 1995; Carranza et al., 2007; Lozupone et al., 2007; Studeny et al., 2011; Gotelli and Chao, 2013; Buckland et al., 2017). For instance, the shape of the curve conveys the evenness of a community. A perfectly even community is represented by a horizontal line (D does not change as q increases) and a highly uneven community is represented by a curve with an initial steep descent as q increases, see https://wagnerbd.shinyapps.io/Frontiers/ (snapshots from the shiny app displayed in Supplementary Figure 1).

Characterization of Diversity Curves Using Bi-Exponential Function

Although visual inspection of diversity curves may identify potential changes in their shape, it is not clear how to make inferences about whether these differences are meaningful. In this section, we propose a method to characterize a sequence of diversity curves using a bi-exponential function.

The D values, alternatively referred to as Hill's numbers (Hill, 1973), are related to the Renyi entropies$\left(H_{(q)}\right)$ (Renyi, 1961) as

$$D_{(q)} = \left(\sum_{k=1}^{K} p_k{}^q\right)^{\frac{1}{1-q}} = e^{H_{(q)}}$$

where Renyi entropies are $H_{(q)} = \frac{1}{1-q} \ln\left(\sum_{k=1}^{K} p_k{}^q\right)$

Suppose taxa can be divided roughly into two groups, rare and non-rare, based on abundance p, and let $k = 1, ..K_1$ for rare taxa with abundance p_1 and $k = K_1 + 1, .., K$ for non-rare taxa with abundance p_2. Then

$$D_{(q)} = \left(\sum_{k=1}^{K} p_k{}^q\right)^{\frac{1}{1-q}},$$
$$\approx \left(K_1 p_1^q + K_2 p_2^q\right)^{\frac{1}{1-q}}$$

where $K_1 + K_2 = K$ and since $e^{\ln(x)} = x$

$$\approx \left(K_1 e^{q*\ln(p_1)} + K_2 e^{q*\ln(p_2)}\right)^{\frac{1}{1-q}}$$

which is now in the form of a bi-exponential function. We can re-parameterize such that

$$\theta_1 = -\ln(p_1),$$
$$\theta_2 = -\ln(p_2),$$
$$\theta_3 = \frac{K_1}{K_1 + K_2}, and$$
$$\theta_4 = K_1 + K_2 \; then$$
$$D_{(q)} = \left(\theta_3\theta_4 e^{q\theta_1} + (1 - \theta_3)\theta_4 e^{q\theta_2}\right)^{\frac{1}{1-q}}$$

where θ_4 is the total number of taxa in the sample, θ_3 is the proportion of rare taxa with a fast rate of decline θ_1 for increasing q and θ_2 is the slow rate of decline in the curve for the $1 - \theta_3$ proportion of non-rare taxa.

TABLE 1 | Comparison of common alpha diversity measures at three time points: Beg Pex, End Pex, and a follow-up visit post-Pex.

		Species observed	Shannon Evenness	Shannon	Simpson Evenness	Simpson
Means (SE)	Beg Pex	26.0 (1.3)	0.46 (0.02)	2.13 (0.12)	0.026 (0.002)	0.62 (0.03)
	End Pex	19.0 (1.3)	0.40 (0.02)	1.70 (0.12)	0.029 (0.002)	0.53 (0.03)
	post-Pex	23.1 (1.4)	0.46 (0.02)	2.07 (0.12)	0.028 (0.002)	0.62 (0.03)
P-values	Across all times	**<0.01**	0.09	**0.01**	0.40	0.07
	Beg vs End	**<0.01**	0.06	**<0.01**	0.19	**0.04**
	Beg vs post-Pex	0.10	0.89	0.69	0.27	0.90
	End vs post-Pex	**0.02**	0.05	**0.02**	0.69	0.06

P-values < 0.05 are indicated in bold.

Development of a Hierarchical Model

In order to make inferences in the changing shape of the curves over time, we propose a longitudinal model to simultaneously estimate the parameters describing the change in the diversity curves over time. To further simplify the model, we will replace the θ_4 parameter with the observed number of taxa and drop the 1/(1-q) exponent. Let

$$\theta_{ijm} = \alpha_{jm} + s_{im}$$

$$D_{(q)ij} = K_{ij}\theta_{ij3}e^{q\theta_{ij1}} + K_{ij}(1-\theta_{ij3})e^{q\theta_{ij2}} + e_{ij}$$

where $m = \{1, 2, 3\}$ indexes the θ parameters for the bi-exponential, $i = 1, .., n$ indexes subjects, $j = 1, 2, 3$ indexes time, α_{jm} is the estimated mean for parameter m at time j, $e_{ij} \sim N(0, \sigma^2)$ is a random error, and s_{im} is a random subject intercept,

$$s_{im} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix}\right)$$

Motivating Example for Alpha Diversity Curves

A non-linear mixed model was estimated by maximum likelihood using SAS PROC NLMIXED. Diversity curves were calculated for each sample and are presented graphically for each subject at each time (**Figure 2A**). The fitted curves follow points indicating good fit. All curves have similar shape and show curvature indicating that the bacterial communities of all samples are relatively uneven (a flatter curve would indicate a more even community). The curves appear to exhibit steeper decline from beginning of the Pex (Beg Pex) to the follow-up visit (post Pex) for the majority of subjects.

The mean curves at each time from the longitudinal model indicate that $D(0)$ (species observed, i.e., richness) was highest at the beginning of the pulmonary exacerbation and decreased thereafter (**Figure 2B**). The curve for the end of the pulmonary exacerbation (End Pex) is more uneven (steeper decline) compared to the other two times. The parameter estimates provided in **Table 2** correspond to visual observations related to the change in shape in both the individual curves and the mean curves, but in addition provide quantification and the ability to make inference on the change in the shape of the curves over time. Estimated θ_1 at End Pex is largest, corresponding to the visually steepest decline, θ_2 estimates increase over time resulting in lower diversity at the Post Pex time associated with the more dominant taxa, and θ_3 estimates from the hierarchical model indicate a significant shift toward a lower proportion of rare taxa over time (**Table 2**).

The shapes of the diversity curves differ (**Figure 2B**) which explains the discrepancies in comparing the diversity indices that were observed earlier (**Table 1**). In addition, using the individual indices conveys no information about evenness without calculating an evenness measure separately. Although separate models evaluating changes in diversity, evenness and richness are easily obtained, there is a different evenness measure corresponding to each diversity measure and therefore this approach suffers from the same issue of multiple measures that may provide different answers depending on how much weight is given to rare taxa. Advantageously, diversity curves provide information about the change in the evenness of the communities over time in a single model. These characteristics can be evaluated and compared numerically with the longitudinal model that allows estimation of trends in the four parameters from the bi-exponential distribution and additional estimates of non-linear functions of these parameters. Application of the hierarchical model to the parameters from the bi-exponential distribution represents a novel approach to evaluating changes in the diversity curves over time.

BETA DIVERSITY

In addition to partitioning diversity into independent components describing evenness and richness, we can also partition diversity by collections of samples. Whereas, the diversity associated with a single sample is referred to as a local (alpha) component, the diversity of the collection of samples is referred to as the regional (gamma) component and the relationship between these two is referred to as beta diversity (Legendre and Legendre, 1998). Previously, an alpha diversity measure was calculated for each sample $\alpha(x_{ij})$, here, a beta diversity index is calculated for each pair of samples $\beta(x_{ij}, x_{ij'})$, and represents either a similarity or a distance between the two samples.

Changes in alpha diversity over time can be useful for evaluating the change in the community structure over time as

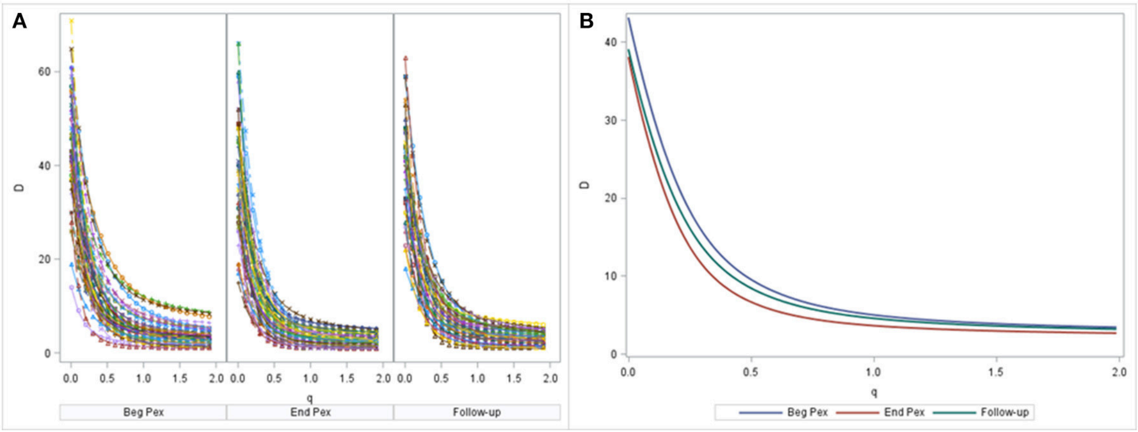


FIGURE 2 | Diversity curves from each sample, where the points correspond to D values from the Hill's numbers (y-axis) plotted vs. the *q* values (x-axis). The corresponding bi-exponential distribution fits are displayed using lines for each time point separately **(A)**. The average diversity curves at each time estimated from the joint longitudinal model are displayed in panel **(B)**.

TABLE 2 | Parameter estimates from nonlinear mixed model at three time points: Beg Pex, End Pex, and a follow-up visit post-Pex.

Est (95% CI)	Beg Pex	End Pex	post-Pex
θ_1	3.65 (2.97–4.33)	3.87 (3.19–4.55)	3.70 (3.02–4.38)
θ_2	1.48 (0.90–2.06)	1.63 (1.05–2.21)	1.64 (1.06–2.22)
θ_3	0.82 (0.72–0.91)	0.79 (0.69–0.90)	0.62 (0.47–0.77)

previously discussed. However, these measures do not convey any information about changes in the community composition (Yuan et al., 2016; Buckland et al., 2017), for example, a community can experience a complete shift in composition, where no taxa are shared, but can still have similar alpha diversity measures, i.e., similar number and abundance of taxa. An important addition to evaluating a microbial community over time in any longitudinal analysis is the incorporation of beta diversity.

As with the alpha diversity measures, there are several possible beta diversity indices that one could use, some of the most popular in microbiome studies include Jaccard, Bray-Curtis, Morisita-Horn and Sorenson. Similar to the earlier discussion of alpha diversity measures, differing results are obtained across beta diversity indices, again due to differences in weighting of the components (Tuomisto, 2010a,b). The calculation of beta diversity indices for all combinations of pairs of samples results in a distance matrix that is often used for ordination (e.g., principal coordinates analysis) and data exploration in microbiota data analysis. Several methods are available for analysis of the full distance matrix (correspondence analysis, redundancy analysis, Mantel test, etc.) (Tuomisto and Ruokolainen, 2006). We focus here on regression based methods that allow for inference at the subject level in a longitudinal design, i.e., studying changes over time within a subject. The implication of this focus is that not all values in the distance matrix are of interest, only those that are comparisons of samples collected within a subject.

Pairwise Comparison of Consecutively Collected Samples

In order to evaluate beta diversity indices at the subject level and compare values over time or across groups, specific values from the full distance or similarity matrix are selected for analysis. In the case of longitudinal studies, we are most interested in evaluating changes in the community over time within a subject and can therefore select the distance measures between samples collected on the same subject $\beta(x_{ij}, x_{ij'})$. One approach that has been used is to calculate the mean or median beta diversity value for each subject and use this as an outcome (Gajer et al., 2012). Here we instead use the beta diversity values from consecutively collected samples within the same subject $\beta(x_{ij}, x_{ij+1})$ as outcomes in a second stage generalized linear mixed model.

Community Turnover

A recently proposed approach in the ecological literature is to use beta diversity indices to evaluate temporal turnover (Collins et al., 2000; Shimadzu et al., 2015). Here, the beta diversity indices are regressed on a time lag variable using a time series model. With this approach, all pairwise indices comparing samples within a subject are used $\beta(x_{ij}, x_{ij'})$ (not simply the indices from consecutive samples as above) and a rate of change in composition is estimated. The proposed approach has been useful for assessing turnover in a single community over time (Collins et al., 2000; Korhonen et al., 2010; Yuan et al., 2016; Lewthwaite et al., 2017), but requires extension to a hierarchical model to make inferences on groups of communities (i.e., subjects in our motivating example). We suggest the use of a similar model to that used for the indices of consecutive samples and simply replace the single independent time variable with one denoting all pairs (Wagner et al., 2017).

Shannon Beta

Another useful measure that has been proposed in the ecological literature (Marcon et al., 2012, 2014) and applied to microbiota data (Zemanick et al., 2015) is the Shannon Beta index. This measure can be decomposed into multiple alpha and beta components even when community weights are unequal (Tuomisto, 2010a,b; Marcon et al., 2012). Thus, in addition to being widely used in other disciplines, its well-understood mathematical properties and underlying theory make Shannon Beta a useful measure overall.

This approach extends the beta diversity measure to apply to a collection of samples rather than just for pair-wise comparisons $\beta(x_{ij}, x_{ij'}, x_{ij''}, ..)$. For our example, Shannon Beta (H_{β_i}) is calculated as

$$H_{\beta_i} = \sum_j \frac{c_{ij}}{c_{i++}} \sum_k \frac{c_{ijk}}{c_{ij+}}\, ln \left(\frac{\frac{c_{ijk}}{c_{ij+}}}{\frac{c_{i+k}}{c_{+++}}} \right)$$

where c_{ijk} is the sequence count for subject i, from time j and taxon k, the $+$ in the subscript denotes the summation of the counts over the specified indicator.

For ease of clinical interpretation, Shannon Beta is expressed as a Hill's number which indicates the effective number of communities represented by the collection of samples or the number of distinct communities. This measure is dependent on the number of samples from which it was calculated, and ranges from 1 to 3 in our motivating example. A normalizing transformation was used to rescale the Hill's numbers to allow comparison across subjects with differences in the number of collected samples (Chao et al., 2010).

$$H_{ni} = \frac{H_{\beta_i} - 1}{j_i - 1}$$

where j_i is the number of samples for subject i.

Motivating Example for Beta Diversity

Morisita-Horn (MH, Beta-diversity) values for pairwise samples j and j' within each subject i were calculated as follows

$$Morisita\ Horn(x_{ij}, x_{ij'}) = 2 \left(\frac{\sum_k \left(c_{ijk} * c_{ij'k}\right)}{\left(\frac{\sum_k {c_{ijk}}^2}{{c_{ij+}}^2} + \frac{\sum_k {c_{ij'k}}^2}{{c_{ij'+}}^2} \right) \left(c_{ij'+} * c_{ij'+}\right)} \right)$$

MH was compared over time using a log-normal model and included a random subject effect. MH in this example is a similarity measure bound between 0 and 1. Values closer to 1 indicate the pair of samples are more similar. MH values, on average, are similar for the two consecutive sample pairs (Beg vs. End Pex and End Pex vs. Follow-up), but individual subjects have varying patterns (**Figure 3A**). Specifically, there are several subjects with limited similarity between communities (MH for both consecutive pairs is close to 0). The turnover analysis was performed in two ways, first, time was defined using the clinically meaningful states (Beg Pex, End Pex, and Post Pex) and second, time was defined using the number of days between when samples were collected. The latter approach is used for illustrative purposes to better show the differences between the consecutive and turnover analyses for this particular example given the small number of samples collected per subject. The turnover analysis using the clinically meaningful time points (**Figure 3B**) reveals that the bacterial communities at Post Pex are more similar to the communities at the Beg Pex than the other comparisons (in this case the consecutive sample comparisons: Beg vs. End Pex and End Pex vs. Post Pex). This indicates that the communities are converging back to the original communities observed at the beginning of the Pex after being perturbed by antibiotics. This is also evident in those subjects with very different communities between consecutively collected samples but show a much higher degree of similarity between Beg Pex and Post Pex. This same pattern is seen using the continuous version of the time variable, where the average similarity values increase with increasing time lag between pairs up to approximately 45 days, after which the similarity declines over time (**Figure 3C**). These figures also illustrate the large amount of variability across subjects with varying patterns in change over time. For both turnover analyses, there are individual subjects whose communities remain stable (no change in similarity with increasing time lag) and those whose communities indicate a directional change (similarity decreases with increasing time lag). A hierarchical model allows each subject's trajectory to deviate from the overall average, capturing this between subject variability. It may be useful to further evaluate the estimated individual subject trajectories by identifying subjects with specific patterns of change over time or by identifying groups of subjects with similar trajectories.

A single Beta diversity measure, Shannon Beta diversity, was calculated for each subject to quantify the number of bacterial communities represented. The median of the beta diversity values after normalization was 0.15 and ranged from 0.04 to 0.75 (**Figure 3D**). Higher values indicate that more distinct communities were observed for a subject, this value ranges from 0 to 3 (number of samples collected per subject). For the subset of subjects with all three samples collected, the median of the Hill's beta diversity measure was 1.3 and ranged from 1.1 to 2.2 and 50% of the values were between 1.2 and 1.6 indicating that the majority of subjects did not experience large shifts in their bacterial communities across all three time points as the number of distinct communities (i.e., Hill's numbers) were around 1.

Both alpha and beta diversity measures from a single example subject are displayed in **Figure 4**. For this subject, the Shannon diversity ($q = 1$) decreases for the second time point and then increases at the third time point but remains below the values observed at the first sample. The communities are very uneven (include several rare taxa) and the diversity curves cross each other indicating that different measures would yield different results, especially for the second and third samples which would differ with lower q values but show similar community characteristics for larger values. The bar charts display the composition of the three communities and show that despite the second and third sample having similar alpha diversity values, the communities are very different.

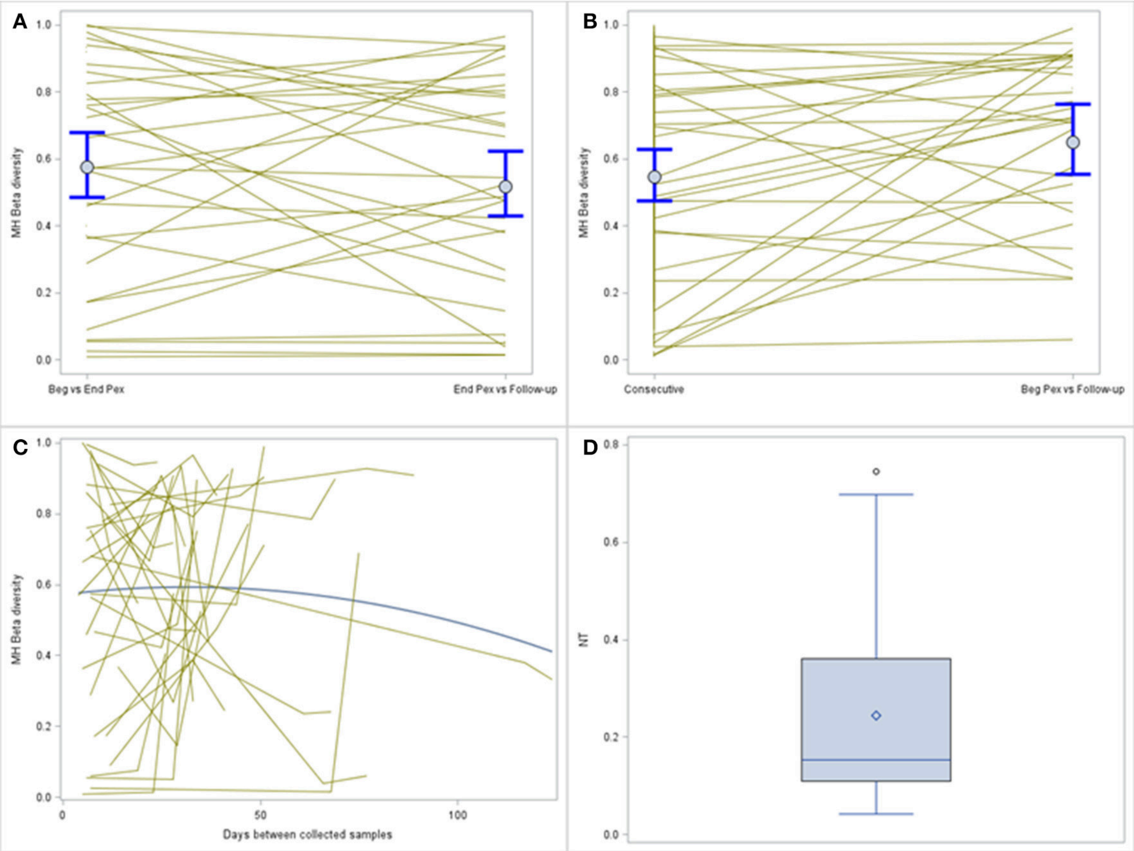


FIGURE 3 | Comparison of MH beta diversity measures for the consecutively collected samples **(A)** and plotted vs. time lag **(B)**. Each subjects value is plotted and connected with lines and the means and 95% confidence intervals from the generalized linear models are plotted with dots and whiskers. The bottom panel displays the MH beta diversity measures plotted over actual time between sample collection **(C)**, individual subjects are indicated by the thin gray lines and the thicker blue line indicates the average change. The distribution of the normalized Shannon Beta diversity measures for all subjects **(D)**.

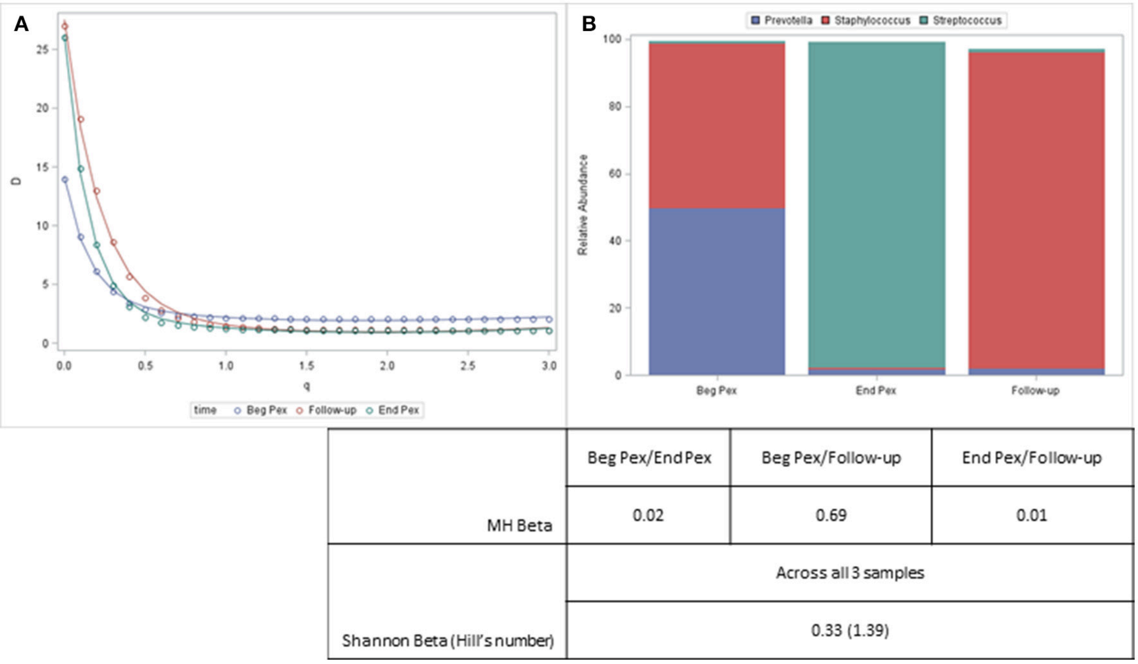
This information, however, is captured with the beta diversity measures. The pairwise MH similarity values illustrate that the samples collected consecutively differ, but that the first sample and the third sample have similar compositions. This indicates that after antibiotics this subject's bacterial community more closely resembled their starting community. The Hill's number for the Shannon beta measure indicates that approximately 1.4 distinct communities are observed for this subject.

The three different approaches to evaluating beta diversity measures in longitudinal studies discussed here provide additional information about changes in communities over time that are not captured by simply modeling alpha diversity over time. The pairwise measures are useful for identifying subjects or times at which shifts in the community are observed and the turnover analysis can yield insights into whether there are consistent shifts with increasing time between sample collections. The single measure (Shannon beta) calculated for each subject can aid in identifying subjects with similar communities across all the time points or those with large changes that suggest shifts over time.

DISCUSSION

In this work, a discussion of methods for evaluating diversity measures in longitudinal microbial data includes the commonly used approach of modeling a single alpha diversity measure over time. Modeling one alpha diversity measure over time (e.g., species observed) could result in different inferences than modeling a different alpha diversity measure (e.g., Simpson). Diversity curves and their calculation were reviewed as a way to alleviate the need to select a single measure; however, until now, there has been no discussion of how to compare curves over time quantitatively. We developed an approach that utilizes a bi-exponential distribution to summarize each curve and compare curves over time using a hierarchical model. This represents a contribution to the field of microbiome data analysis. Lastly, we discuss the additional information that is gained by evaluating beta diversity measures to assess changes in community composition over time and implement three different approaches and discuss their differences.

Several measures of diversity have been widely applied to microbiota data and have been used previously as outcomes



	Beg Pex/End Pex	Beg Pex/Follow-up	End Pex/Follow-up
MH Beta	0.02	0.69	0.01
	Across all 3 samples		
Shannon Beta (Hill's number)	0.33 (1.39)		

FIGURE 4 | Diversity curves for an example subject **(A)** corresponding to the communities represented by the stacked barcharts **(B)**. Taxa with a relative abundance > 5% for any sample are displayed. The table shows the pairwise MH beta diversity values and the Shannon Beta for this subject.

in longitudinal models (Gajer et al., 2012; Flores et al., 2014; Wagner et al., 2017). Often these indices are incorrectly treated as interchangeable measures of the same characteristic, which has caused debates and general confusion over which diversity measure to use, misinterpretation of results, and an inability to compare results across studies. Diversity curves incorporating aspects of these different measures were promoted as a solution, but until now have been used simply for visualization purposes (Renyi, 1961; Whittaker, 1972; Hill, 1973; Carranza et al., 2007; Studeny et al., 2011; Gotelli and Chao, 2013). We chose to model the diversity curves proposed by Jost (2006) that have been shown to equal several commonly used measures, although we recognize that there are alternative complexity curves that have been proposed (Rajaram and Castellani, 2016). The use of any curve will require a model to be applied to capture the shape of the curve to make inferences about changes over time.

An important addition to evaluating a microbial community over time in any longitudinal analysis is the incorporation of beta diversity, as these measures convey information about changes in community composition. The majority of previous analyses have concentrated on modeling turnover in a single community (Collins et al., 2000; Korhonen et al., 2010; Yuan et al., 2016; Lewthwaite et al., 2017). Two studies (Gajer et al., 2012; Wagner et al., 2017) modeled beta diversity measures over time using a hierarchical model similar to the model using beta diversity from consecutive times discussed in this paper, but the descriptions of the models were relegated to supplements. In this paper we describe in detail the modeling approach and its interpretation. Our example included the Morisita-Horn beta diversity measure, selected because it is not influenced by richness and sequencing effort (Magurran, 2004). However, various other beta diversity measures including phylogenetic measures that account for genetic similarity between taxa can be used (Gotelli and Chao, 2013) without loss of generality of the modeling approach.

The Shannon Beta index is another useful measure that has been proposed in the ecological literature (Marcon et al., 2012, 2014) and applied to microbiota data (Zemanick et al., 2015). This measure provides a single number denoting the similarity across multiple communities and can be used to identify subjects with small or large changes in their bacterial community. To our knowledge, the Shannon Beta index has not been previously applied to evaluate changes in bacterial communities within a subject over time and thus our methods represent a novel application of this measure to longitudinal microbiota data.

All of the methods discussed are illustrated and compared using a motivating example in cystic fibrosis. The example included a small number of repeated samples per subject and samples corresponded to clinically meaningful time points (hospital admission, hospital discharge, and a follow-up visit post-exacerbation). For this reason, the models we employed designated time as a categorical variable. These models are flexible and could include time as a continuous variable instead for studies with more longitudinal samples collected. The separate models for alpha diversity indices indicated that diversity decreased with administration of antibiotics mainly driven by a decrease in richness. This pattern was also observed in modeling the alpha diversity curves and provided the ability to make inferences about the components of diversity (richness and evenness) without requiring a separate model for each measure. Alpha diversity can provide information about changes in community structure but does not provide any information about changes in community composition, to address this, beta diversity measures are needed. The majority of studies utilizing beta diversity, use the measures to perform exploratory data analysis with ordination plots. Here, we chose to focus on models of beta diversity that can be used to test hypotheses about change in community composition over time. We illustrated three different approaches for modeling beta diversity. The first used the beta diversity measure from consecutively collected samples, and showed that the average MH was fairly large indicating similar communities. However, there were subjects with a high degree of dissimilarity between consecutive samples (MH values close to 0), whereas the turnover analysis revealed that for these subjects, there were large changes while on treatment but the follow-up community reverted back to the baseline community after being perturbed with antibiotics. Given the small number of samples collected per person in the motivating example, this pattern could have been discerned by evaluating beta diversity for all three combinations of sample pairs, an example with more samples per person or unbalanced collection (samples collected at different times) might have greater benefit from the insight gained from both analyses. The third approach provided a single measure per subject that compares composition of all three samples. This method does not provide information about trends over time but it can be used to rank subjects based on whether they had large changes or whether the three communities were relatively similar. This information could be useful for correlating with clinically important factors, like whether the subject exhibited clinical improvement with treatment.

It was necessary to select specific approaches/indices to include in this work. We recognize though that different alternatives could have been chosen. Instead of providing an exhaustive list and comparison of all methods, we chose approaches that provided good examples of the concepts with the understanding that the methods discussed generalize to other measures; any beta diversity and any measure could be used as the outcome in the models discussed. Future work could incorporate the efforts to classify beta diversity measures based on differences in weighting of the components (Tuomisto, 2010a,b) for application to longitudinal studies.

In summary, several approaches to analyzing diversity measures in a longitudinal study were discussed and compared, including a novel approach modeling longitudinal patterns in alpha diversity curves over time. Given the importance of repeated sampling of microbial communities, especially in human studies, extension of methods appropriate for longitudinal study designs are needed.

AUTHOR CONTRIBUTIONS

BW and JH: Conception and design. JH, CR, and EZ: Acquisition of the data. BW and CR: Performed the analysis. BW, JH and GG: Drafting the manuscript for important intellectual content. BW, JH, GG, GZ, SM-G, CR and EZ: Review and revision of manuscript.

FUNDING

This work was supported by the National Institutes of Health (NIH/NCATS Colorado CTSA Grant Number UL1 TR001082, K23HL114883, 1R01HL124103, R01 DA035804, and R01 DA032555), Children's Hospital Colorado Research Institute and the Cystic Fibrosis Foundation (WAGNER15A0, ZEMANI12A0, ZEMANI16A0).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01037/full#supplementary-material

REFERENCES

Buckland, S. T., Yuan, Y., and Marcon, E. (2017). Measuring temporal trends in biodiversity. *Adv. Stat. Anal.* 101, 461–474. doi: 10.1007/s10182-017-0308-1

Carranza, M. L., Acosta, A., and Ricotta, C. (2007). Analyzing landscape diversity in time: the use of Renyi's generalized entropy function. *Ecol. Indic.* 7, 505–510. doi: 10.1016/j.ecolind.2006.05.005

Chao, A., Chiu, C. H., and Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 3599–3609. doi: 10.1098/rstb.2010.0272

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

Collins, S. L., Micheli, F., and Hartt, L. (2000). A method to determine rates and patterns of variability in ecological communities. *Oikos* 91, 285–293. doi: 10.1034/j.1600-0706.2000.910209.x

Ellison, A. M. (2010). Partitioning diversity. *Ecology* 91, 1962–1963. doi: 10.1890/09-1692.1

Fang, R., Wagner, B. D., Harris, J. K., and Fillon, S. A. (2016). Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiol. Infect.* 144, 2477–2455. doi: 10.1017/S0950268816000662

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol.* 15:531. doi: 10.1186/s13059-014-0531-y

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4:132ra52. doi: 10.1126/scitranslmed.3003605

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

Gotelli, N. J., and Chao, A. (2013). "Measuring and estimating species richness, species diversity, and biotic similarity from sampling data," in *Encyclopedia of Biodiversity*, ed S. A. Levin (Waltham, MA: Academic Press), 195–211.

Hara, N., Alkanani, A. K., Ir, D., Robertson, C. E., Wagner, B. D., Frank, D. N., et al. (2012). Prevention of virus-induced type 1 diabetes with antibiotic therapy. *J. Immunol.* 189, 3805–3814. doi: 10.4049/jimmunol.1201257

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432. doi: 10.2307/1934352

Jost, L. (2006). Entropy and diversity. *Oikos* 113, 363–375. doi: 10.1111/j.2006.0030-1299.14714.x

Jurasinski, G., and Koch, M. (2011). Commentary: do we have a consistent terminology for species diversity? We are on the way. *Oecologia* 167, 893–902, discussion: 903–911. doi: 10.1007/s00442-011-2126-6

Korhonen, J. J., Soininen, J., and Hillebrand, H. (2010). A quantitative analysis of temporal turnover in aquatic species assemblages across ecosystems. *Ecology* 91, 508–517. doi: 10.1890/09-0392.1

Legendre, P., and Legendre, L. (1998). *Numerical Ecology.* Amsterdam; New York, NY: Elsevier.

Lewthwaite, J. M. M., Debinski, D. M., and Kerr, J. T. (2017). High community turnover and dispersal limitation relative to rapid climate change. *Glob. Ecol. Biogeogr.* 26, 459–471. doi: 10.1111/geb.12553

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi: 10.1128/AEM.01996-06

Magurran, A. E. (2004). *Measuring Biological Diversity*, Malden, MA: Blackwell Science Ltd.

Marcon, E., Herault, B., Baraloto, C., and Lang, G. (2012). The decomposition of Shannon's entropy and a confidence interval for beta diversity. *Oikos* 121, 516–522. doi: 10.1111/j.1600-0706.2011.19267.x

Marcon, E., Scotti, I., Hérault, B., Rossi, V., and Lang, G. (2014). Generalization of the partitioning of shannon diversity. *PLoS ONE* 9:e90289. doi: 10.1371/journal.pone.0090289

Markle, J. G., Frank, D. N., Mortin-Toth, S., Robertson, C. E., Feazel, L. M., Rolle-Kampczyk, U., et al. (2013). Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* 339, 1084–1088. doi: 10.1126/science.1233521

Moreno, C. E., and Rodríguez, P. (2011). Commentary: do we have a consistent terminology for species diversity? Back to basics and toward a unifying framework. *Oecologia* 167, 889–892; discussion 903–911. doi: 10.1007/s00442-011-2125-7

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rajaram, R., and Castellani, B. (2016). An entropy based measure for comparing distributions of complexity. *Physica A* 453, 35–43. doi: 10.1016/j.physa.2016.02.007

Renyi, A. (1961). "On measures of entropy and information," in *4th Berkeley Symposium on Mathematical Statistics and Probability,* ed J. Neyman (Berkeley, CA: University of California Press), 547–561.

Robertson, C. E., Harris, J. K., Wagner, B. D., Granger, D., Browne, K., Tatem, B., et al. (2013). Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics* 29, 3100–3101. doi: 10.1093/bioinformatics/btt526

Shimadzu, H., Dornelas, M., and Magurran, A. E. (2015). Measuring temporal turnover in ecological communities. *Methods Ecol. Evol.* 6, 1384–1394. doi: 10.1111/2041-210X.12438

Studeny, A. C., Buckland, S. T., Illian, J. B., Johnston, A., and Magurran, A. E. (2011). Goodness-of-fit measures of evenness: a new tool for exploring changes in community structure. *Ecosphere* 2, 1–19. doi: 10.1890/ES10-00074.1

Tothmeresz, B. (1995). Comparison of different methods for diversity ordering. *J. Veg. Sci.* 6, 283–290. doi: 10.2307/3236223

Tuomisto, H. (2010a). A diversity of beta diversities: straightening up a concept gone awry. Part 2. quantifying beta diversity and related phenomena. *Ecography* 33, 23–45. doi: 10.1111/j.1600-0587.2009.06148.x

Tuomisto, H. (2010b). A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography* 33, 2–22. doi: 10.1111/j.1600-0587.2009.05880.x

Tuomisto, H. (2011). Commentary: do we have a consistent terminology for species diversity? Yes, if we choose to use it. *Oecologia* 167, 903–911. doi: 10.1007/s00442-011-2128-4

Tuomisto, H., and Ruokolainen, K. (2006). Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology* 87, 2697–2708. doi: 10.1890/0012-9658(2006)87[2697:AOEBDU]2.0.CO;2

Wagner, B. D., Sontag, M. K., Harris, J. K., Miller, J. I., Morrow, L., Robertson, C. E., et al. (2017). Airway microbial community turnover differs by BPD severity in ventilated preterm infants. *PLoS ONE* 12:e0170120. doi: 10.1371/journal.pone.0170120

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon* 21, 213–251. doi: 10.2307/1218190

Yuan, Y., Buckland, S. T., Harrison, P. J., Foss, S., and Johnston, A. (2016). Using species proportions to quantify turnover in biodiversity. *J. Agric. Biol. Environ. Stat.* 21, 363–381. doi: 10.1007/s13253-015-0243-0

Zemanick, E. T., Wagner, B. D., Robertson, C. E., Ahrens, R. C., Chmeil, J. F., Clancy, J. P., et al. (2017). Airway microbiota across age disease spectrum in cystic fibrosis. *Eur. Respir. J.* 50:1700832. doi: 10.1183/13993003.00832-2017

Zemanick, E. T., Wagner, B. D., Robertson, C. E., Stevens, M. J., Szefler, S. J., Accurso, F. J., et al. (2015). Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. *Ann. Am. Thorac. Soc.* 12, 221–229. doi: 10.1513/AnnalsATS.201407-310OC

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 Microbiology

METHODS
published: 26 July 2018
doi: 10.3389/fmicb.2018.01683

Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data

Xinyan Zhang[1†], *Yu-Fang Pei*[2†], *Lei Zhang*[2], *Boyi Guo*[3], *Amanda H. Pendegraft*[3], *Wenzhuo Zhuang*[4] *and Nengjun Yi*[3*]

[1] Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, United States, [2] Department of Epidemiology and Health Statistics, School of Public Health, Medical College of Soochow University, Suzhou, China, [3] Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States, [4] Department of Cell Biology, School of Biology & Basic Medical Science, Soochow University, Suzhou, China

The metagenomics sequencing data provide valuable resources for investigating the associations between the microbiome and host environmental/clinical factors and the dynamic changes of microbial abundance over time. The distinct properties of microbiome measurements include varied total sequence reads across samples, over-dispersion and zero-inflation. Additionally, microbiome studies usually collect samples longitudinally, which introduces time-dependent and correlation structures among the samples and thus further complicates the analysis and interpretation of microbiome count data. In this article, we propose negative binomial mixed models (NBMMs) for longitudinal microbiome studies. The proposed NBMMs can efficiently handle over-dispersion and varying total reads, and can account for the dynamic trend and correlation among longitudinal samples. We develop an efficient and stable algorithm to fit the NBMMs. We evaluate and demonstrate the NBMMs method via extensive simulation studies and application to a longitudinal microbiome data. The results show that the proposed method has desirable properties and outperform the previously used methods in terms of flexible framework for modeling correlation structures and detecting dynamic effects. We have developed an R package NBZIMM to implement the proposed method, which is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM and provides a useful tool for analyzing longitudinal microbiome data.

Keywords: count data, longitudinal study, microbiome, metagenomics, negative binomial mixed model

OPEN ACCESS

Edited by:
Michele Guindani,
University of California, Irvine,
United States

Reviewed by:
Ye Deng,
Research Center for
Eco-Environmental Sciences (CAS),
China
Qing Lu,
Michigan State University,
United States

***Correspondence:**
Nengjun Yi
nyi@uab.edu

[†] These authors have contributed equally to this work.

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

Received: *27 October 2017*
Accepted: *06 July 2018*
Published: *26 July 2018*

Citation:
Zhang X, Pei Y-F, Zhang L, Guo B, Pendegraft AH, Zhuang W and Yi N (2018) Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. Front. Microbiol. 9:1683. doi: 10.3389/fmicb.2018.01683

INTRODUCTION

The human microbiome plays an important role in human health and disease. The complex microbiome is inherently dynamic and interacts with the host and the environmental factors over time (Gerber, 2014a). These complex dynamics start from the birth with increasingly richness in the communities of microbiota over time (Palmer et al., 2007; Koenig et al., 2011; Wu et al., 2011; De Muinck et al., 2013; Gerber, 2014a). Recent studies have found that the human microbiome in healthy adults can be altered by various host factors including genotype (Spor et al., 2011; Blekhman et al., 2015; Goodrich et al., 2016a,b), lifestyle such as dietary habit (De Filippo et al., 2010; Wu et al., 2011), physiological status such as aging (Biagi et al., 2010), pathophysiological status (Turnbaugh et al., 2009), and host environment (Dominguez-Bello et al., 2010). The dynamic shifts in compositional features of the microbiome can occur with human diseases such as obesity (Turnbaugh et al., 2006), diabetes (Samuel and Gordon, 2006), infections or inflammatory bowel disease (Frank et al., 2007), and cancers (Holmes et al., 2011). To decipher the relationship between the dynamic changes in microbiome and human diseases, high-throughput sequencing technologies, such as the 16S ribosome RNA (rRNA) gene sequencing or shotgun metagenomics sequencing, have been widely applied in longitudinal microbiome studies (Matsen et al., 2010; Ghodsi et al., 2011; Gilbert et al., 2011; La Rosa et al., 2014).

The metagenomics sequencing data provide valuable resources for investigating the dynamic changes of microbial abundance over time and the associations between the microbiome and host environmental/clinical factors. Multiple recent microbiome studies have employed the longitudinal study designs to address the crucial research question (La Rosa et al., 2014; DiGiulio et al., 2015; Zhou et al., 2015; Ward et al., 2016). Among them, La Rosa et al. (2014) utilized longitudinal analysis of repeated measures data to demonstrate that the dynamic shifts in dominating microbiota of the infant gut from *Bacilli* at birth, giving way to *Gammaproteobacteria*, then *Clostridia* at the end of the first month of life. In another recent published study, Ward et al. (2016) used longitudinal study to address the associations between the dynamic change of the early intestinal microbiome in preterm infants and the occurrence of Necrotizing enterocolitis (NEC) or NEC-associated deaths.

Despite our ability to generate large-scale metagenomics sequencing longitudinal data, many challenges exist in the development of robust and powerful statistical methods and computational tools for properly analyzing and interpreting longitudinal microbiome data. The metagenomics sequencing data has some properties that require tailored analytic tools; these include varied total sequence reads across samples, over-dispersion and zero-inflation. One common way to account for varying total reads is normalization, i.e., conversion of the sequence counts to the relative abundance (or proportion) using the total sum, mean, or median of representative OTUs across all samples (Anders and Huber, 2010; Robinson and Oshlack, 2010; Knights et al., 2011; Wagner et al., 2011; Kostic et al., 2012; Paulson et al., 2013). Several zero-inflated models were proposed to correct for excess zero counts in microbiome measurements, including zero-inflated Gaussian, lognormal, negative binomial, and beta models (Paulson et al., 2013; Peng et al., 2015; Sohn et al., 2015; Xu et al., 2015). On the other hand, the negative binomial regression, which is a standard statistical method for analyzing over-dispersed count observations, has been recently applied to microbiome data (White et al., 2009; Pookhao et al., 2015).

It is even more challenging to analyze longitudinal microbiome count data. In addition to the special features of microbiome data, longitudinal studies possesses two fundamental time-dependent features: (a) time imposes an inherent and irreversible ordering on samples, and (b) samples exhibit statistical dependencies that are a function of time (Gerber, 2014b). Ignoring these properties of longitudinal data and applying statistical tools designed for analyzing static data can result in erroneous conclusions (Gerber, 2014a). Most of the previous studies resort to linear mixed models (LMMs) to account for time-dependent correlations in longitudinal microbiome study designs by treating transformed data as normally distributed responses (Benson et al., 2010; Srinivas et al., 2013; La Rosa et al., 2014; Leamy et al., 2014; Wang et al., 2015). However, using LMMs directly without addressing properties of microbiome data may result in lower power or potential inaccurate results to detect the dynamic effects of microbiota. Chen and Li (2016) developed zero-inflated beta mixed models for analyzing transformed proportions in microbiome longitudinal studies, but did not address time trends and within-subject correlations. Thus, statistical models to account for time series as well as properties of microbiome count data are required for analyzing microbiome data (Spor et al., 2011; Faust et al., 2015; Chen and Li, 2016).

Zhang et al. (2017) have recently developed negative binomial mixed models (NBMMs) for analyzing clustered microbiome data, but have not addressed longitudinal studies yet. We here extend negative binomial mixed models (NBMMs) proposed by Zhang et al. (2017) to analyze longitudinal microbiome count data. The extended NBMMs can include various types of fixed effects and random effects, and can incorporate various correlation structures among observations within the same subjects, thus fully addressing the special properties of longitudinal microbiome count data. We develop an efficient and stable IWLS (iterative weighted least squares) algorithm to fit the extended NBMMs by taking advantage of the standard procedure for fitting linear mixed models. Through extensive simulations, we show that the NBMMs outperform the previously used LMMs in terms of detecting dynamic effects in longitudinal microbiome count data. We also apply our method to a previously published microbiome data to detect significantly dynamic effects of associated taxa. We have implemented the proposed method in the R package NBZIMM, which is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM and provides a useful tool for longitudinal microbiome studies.

METHODS

Negative Binomial Mixed Models (NBMMS) for Longitudinal Microbiome Studies

Longitudinal studies collect multiple subjects and measure each subject at multiple time points (i.e., samples). Assume that there are n subjects, and subject i is measured at n_i time points t_{ij}; $j = 1, \cdots, n_i$; $i = 1, \cdots, n$. For each sample, microbiome data generated by the 16S rRNA gene sequencing or the shotgun metagenomics sequencing consist of counts for numerous taxa at certain taxonomic levels (OTU, species, genus, classes, etc.), c_{ijh}, $h = 1, \cdots, m$, and total sequence read T_{ij} (also referred to as depths of coverage or library size). We also measure some host clinical/environmental variables for each subject, X_i. **Table 1** summarizes the data structure for a longitudinal microbiome study. The goal of longitudinal microbiome studies is to detect associations between the microbiome counts and the

TABLE 1 | Longitudinal microbiome data structure.

Subject ID	Taxon 1	Taxon 2	···	Taxon *m*	Total reads	Host factors	Time variables
Subject 1	c_{111}	c_{112}	···	c_{11m}	T_{11}	X_1	t_{11}
Subject 1	c_{121}	c_{122}	···	c_{12m}	T_{12}	X_1	t_{12}
Subject 1	c_{131}	c_{132}	···	c_{13m}	T_{13}	X_1	t_{13}
Subject 2	c_{211}	c_{212}	···	c_{21m}	T_{21}	X_2	t_{21}
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
Subject n	c_{n11}	c_{n12}	···	c_{n1m}	T_{n1}	X_n	t_{n1}

host variables, and characterize the time trends of microbiome abundance within subjects and between subjects.

We separately analyze each microbiome taxon, as most existing methods. For notational simplification, we denote $y_{ij} = c_{ijh}$ for any given taxon h. Since the microbiome count outcome is over-dispersed, we use negative binomial models. We extend negative binomial mixed models (NBMMs) proposed by Zhang et al. (2017) to analyze longitudinal microbiome data by including the time variable and its interaction with the host factor of interest in the model. In the next section, we will further extend NBMMs to account for within-subject correlation structures.

In our NBMMs, the counts y_{ij} are assumed to follow the negative binomial distribution:

$$y_{ij} \sim \ NB(y_{ij} \mid \mu_{ij}, \theta) = \frac{\Gamma(y_{ij} + \theta)}{\Gamma(\theta) y_{ij}!} \cdot \left(\frac{\theta}{\mu_{ij} + \theta}\right)^{\theta} \cdot \left(\frac{\mu_{ij}}{\mu_{ij} + \theta}\right)^{y_{ij}} \quad (1)$$

where θ is the dispersion parameter that controls the amount of over-dispersion, and μ_{ij} are the means. The means μ_{ij} are related to the host variables via the logarithm link function:

$$\log(\mu_{ij}) = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_i \quad (2)$$

where $\log(T_{ij})$ is the offset that corrects for the variation of the total sequence reads, $X_{ij} = (1, X_i, t_{ij}, X_i^s t_{ij})$, X_i^s is the variable of interest in X_i, for example, an indicator variable for the case group and the control group, and$Z_{ij} = (1, t_{ij})$; $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is the vector of fixed effects (i.e., population-level effects), including an intercept β_0, the effects β_1 of the host variables X_i, the overall time effect β_2, and the interaction β_3 betweenX_i^s andt_{ij}; $b_i = (b_{0i}, b_{1i})^{\mathrm{T}}$ is the vector of random effects (i.e., subject-level effects), including the random intercept b_{0i} and the random time effect b_{1i}. For simplicity, the above model only considers the linear function of t_{ij}. If sample size is large enough, however, we can extend the model to use polynomial functions, for example, $(t_{ij}, {t_{ij}}^2)$, or B-spline functions, allowing us to detect arbitrary temporal trends.

The random effects are used to model multiple sources of variations and subject-specific effects, and thus avoid biased inference on the fixed effects. The vector of the random effects is usually assumed to follow a multivariate normal distribution (Pinheiro and Bates, 2000; McCulloch and Searle, 2001):

$$b_i \sim \ N(0, \Psi) \quad (3)$$

where Ψ is the variance-covariance matrix. Ψ can be a general positive-definite matrix that accounts for the correlation of the random covariates. In some applications, however, we can restrict Ψ to special forms of variance-covariance matrices that are parameterized by fewer parameters. For example, we may assume that the random effects are independent, in which case Ψ is a diagonal matrix.

Accounting for Within-Subject Correlations and IWLS Algorithm for Fitting the NBMMS

The IWLS (Iterative Weighted Least Squares) algorithm developed by Zhang et al. (2017) can be used to fit the above NBMMs. The basic idea of the IWLS algorithm is to iteratively approximate the negative binomial mixed model by a linear mixed model. However, Zhang et al. (2017) restricts the within-subject errors in the linear mixed model to be independent, and thus ignores special within-subject correlation structures. For longitudinal data, however, samples within the same subject are usually correlated. Thus, we extend the model by relaxing the assumption of independent within-subject errors to account for special within-subject correlation structures:

$$\begin{aligned} z_{ij} = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_i + w_{ij}^{-1/2} e_{ij}, \ b_i \sim \ N(0, \Psi), \\ e_i = (e_{i1}, \cdots, e_{in_i})' \sim \ N(0, \sigma^2 R_i) \quad (4) \end{aligned}$$

where z_{ij} and w_{ij} are the pseudo-responses and the pseudo-weights, respectively, that depend on $\log(T_{ij}) + X_{ij}\hat{\beta} + Z_{ij}\hat{b}_i$ and $\hat{\theta}$as described in Zhang et al. (2017), and R_i is a correlation matrix, which describes dependence among observations, Pinheiro and Bates (2000) describes several ways to specify the correlation matrix R_i, all of which can be incorporated into our NBMMs. For longitudinal studies, a common choice of R_i is autoregressive of order 1, AR(1), or continuous-time AR(1).

We extend the IWLS algorithm developed by Zhang et al. (2017) to fit the proposed NBMMS with correlation structures. The algorithm alternatively updates the dispersion θ and the parameters in the linear mixed model (4). Given the estimates of β and b, we update the dispersion parameter θ by maximizing the negative binomial likelihood using the standard Newton-Raphson algorithm, and then calculate the pseudo-responses and the pseudo-weights. We then fit the linear mixed model (4) using the standard method as implemented in the core package **nmle** in R. At convergence of the algorithm, we get the maximum likelihood estimates of all the fixed effects β_k and their confidence intervals from the final linear mixed model. We then can test H_0: $\beta_k = 0$ following the linear mixed model framework.

R Package for Implementing the Proposed Method

We have created the function **glmm.nb** for setting up and fitting the proposed NBMMs, which is part of the R package **NBZIMM**. The function **glmm.nb** works by repeated calls to the function **lme** for fitting linear mixed models in the recommended package **nlme** in R, and allows for any types of random effects and within-subject correlation structures as described in the package **nlme**. The outputs from the function **glmm.nb** can be summarized by functions in **nlme**. The package **NBZIMM** is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM.

RESULTS

Simulation Studies

Simulation Designs

We performed extensive simulations to evaluate the proposed methods. We extended the simulation framework of Zhang et al. (2017) to simulate longitudinal microbiome counts from negative binomial distributions and incorporate time covariates, random effects and within-subject correlation structures.

Our simulation studies employed a case-control longitudinal study design with four different settings. All the four simulation settings followed a two-level longitudinal study, where all individuals (subjects) were from two groups (i.e., case or control) and multiple samples were measured at several time points for each individual. For all the settings, we simulated ($n =$) 50, 100 or 150 individuals, half of which were cases, and included three fixed covariates: a binary case-control indicator variable x_i, a continuous time variable t_{ij}, and their interaction. We denote the fixed effects of these three covariates by (β_1, β_2, β_3). The time points, random effects, and within-subject correlation structures were set as follows:

1) Setting A: 5 time points for each individual, only random intercept, and no within-subject correlation;
2) Setting B: 10 time points for each individual, only random intercept, and the within-subject correlation was autoregressive of order 1, AR(1);
3) Setting C: 5 time points for each individual, two random effects (i.e., random intercept and time effect), and no within-subject correlation;
4) Setting D: 4 or 5 different time points for individuals, only random intercept, and no within-subject correlation;

To minimize possible bias and yield reasonable count values that are similar to real microbiome data, we randomly generated the parameters in the model from reasonable ranges at each simulation replication (Zhang et al. 2017), which are described as follows:

1) The values, $\log(T_{ij}) + \beta_0$, control the means of simulated counts when all the effects are zero, where β_0 is the fixed intercept. We set $\beta_0 = -7$ and randomly sampled $\log(T_{ij})$ from the range [7.1, 10.5]. In this case, $\log(T_{ij}) + \beta_0$ were in the range [0.1, 3.5], which yield counts similar to real microbiome data;
2) The dispersion parameter θ were uniformly sampled from the range [0.1, 5], which yield highly or moderate over-dispersed counts;
3) To evaluate false positive rates, the fixed effects β_1, β_2 and β_3 were all set to be zero. To evaluate empirical powers, we considered four scenarios: a) β_1 and β_2 were set to 0, and β_3 was sampled from [0.2, 0.35]; b) β_1 and β_2 were set to 0, and β_3 was sampled from [0.35, 0.8]; c) β_1, β_2 and β_3 were all sampled from [0.2, 0.35]; d) β_1, β_2 and β_3 were all sampled from [0.35, 0.8];
4) The random effects b_{0i} and b_{1i} were generated from N(0, τ^2), where τ was randomly drawn from the range [0.5, 1];
5) The correlation coefficient ρ for AR(1) correlation was sampled from [0.1, 0.5], and the AR(1) correlation was generated by the function *arima.sim()* from R package *stats*;
6) The standard deviation σ was sampled from [0.1, 0.5];

TABLE 2 | Parameter ranges in simulation studies.

Parameter	Range
$\log(T_{ij}) + \beta_0$	Unif(0.1, 3.5)
Dispersion parameter θ	Unif(0.1, 5)
Fixed effects β_1, β_2, β_3 (false positive rate)	0, 0, 0
Fixed effects β_1, β_2, β_3 (power of interaction)	0, 0, Unif(0.2, 0.35) or Unif(0.35, 0.8)
Fixed effects β_1, β_2, β_3 (power of both β_1 and β_3)	All from Unif(0.2, 0.35) or Unif(0.35, 0.8)
Standard deviation τ	Unif(0.5, 1)
Correlation ρ	Unif(0.1, 0.5)
Standard deviation σ	Unif(0.1, 0.5)

The ranges of all the parameters used in the simulation are summarized in **Table 2**.

In all the four simulation settings, the procedure was repeated 10,000 times. At each replication, the parameters were sampled from the ranges described above. There were two hypotheses of interests to be tested, i.e., the group main effect $\beta_1 = 0$ and the group by time interaction $\beta_3 = 0$. Both empirical power and false positive rate for testing the hypotheses were calculated under significance level at 0.05. The empirical power and false positive rate were defined as the proportions of detecting non-zero and zero effects over the simulation replications, respectively. We compared the proposed NBMMs with the linear mixed model with the arcsine square root transformation, $arcsine\left(\sqrt{y_{ij}/T_{ij}}\right)$, as the response, denoted by LMM arcsin.

Simulation Results

Figure 1 and Figure A.1 show the empirical power to detect the group by time interaction under the four different simulation settings, when the group main effect was set to zero. The power was affected by the sample size. It can be clearly seen that the proposed method performed consistently better than the LMM arcsin method across almost all the scenarios. The second setting was set to represent time-series structure in longitudinal data with 10 measurements for each individual, and thus had the largest power among all the four settings. It was shown that the first setting had higher power than the third setting, on the other hand, a similar performance in power compared with the fourth setting.

It is of interest to detect both the group main effect and the group by time interaction. Therefore, in another set of parameter settings, we targeted to detect both the group main effect and the group by time interaction. **Figure 2** and Figure A.2 show the

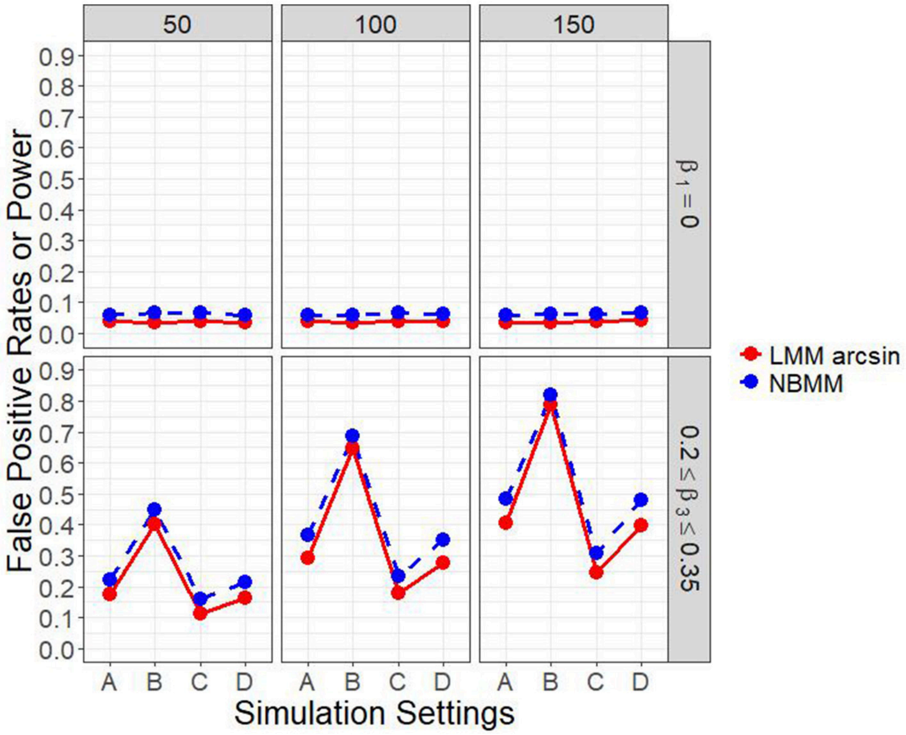


FIGURE 1 | Empirical power of interaction term and false positive rates of main effect in all four simulation settings.

empirical power to detect both the group main effect and the group by time interaction under the four different simulation settings. The results showed that the LMM arcsine method resulted in a slightly higher power in detecting interaction term than our proposed method across all the scenarios. However, it showed an extreme low power close to alpha level in detecting the group main effect across all the scenarios. It inferred that LMM arcsine method is not an appropriate approach to be used when the group main effect and the group by time interaction effect are both nonzero. **Figure 3** displays the false positive rates for detecting both the group main effect and interaction effect. For all the four simulation settings, the false positive rates were well controlled under all the scenarios.

Application to Temporal and Spatial Pregnant Data

We applied our method to a public microbiome data from a longitudinal study to investigate the bacterial taxonomic composition for pregnant and postpartum women by DiGiulio et al. (2015). This case-control longitudinal study included 49 pregnant women, 15 of whom delivered preterm. The discovery data was consisted with 40 of those women. Among those 40 women, they collected 3,767 specimens prospectively and weekly during gestation and monthly after delivery from the vagina, distal gut, saliva, and tooth/gum. The specimens were analyzed for bacterial taxonomic composition. The final dataset contained a total of 1271 taxa from 3432 specimens which were identified for pregnant women delivered at term and preterm. Detailed information about population and material is available in DiGiulio et al. (2015). Clinical data included race, weeks/days when the samples were obtained, way of delivery, and household income level were acquired. The public processed OTU data available from the study is from species level. The clinical data for the validation dataset for the rest of 9 pregnant women is not available.

We used the proposed NBMMs and the linear mixed models (LMMs) with the arcsine square root transformations to detect associations between delivery term and vaginal bacteria taxa composition during pregnancy. The host factor in the analysis was defined as two groups with patients who delivered at preterm vs. term. The patients who delivered at marginal term were excluded from the analysis. Only specimens collected in vaginal during pregnancy were included in the analysis. Meanwhile, according to the original paper, the samples could be divided to 5 Vaginal Community State Types. Only samples with community state type 4 were analyzed in the original paper. To be consistent, we followed the same criteria for sample filtering. The sample size in the final analysis was 103. We included 58 taxa with zero proportion greater than 0.25 for 103 samples in our analysis. The real data and the R code for our analysis are available from the GitHub page: https://abbyyan3.github.io//NBZIMM-tutorial/NBZIMM_NBMMs_Longitudinal.html.

To compare the abilities of LMMs and NBMMs in detecting the static and dynamic association between host factor and vaginal bacterial taxa composition, we used the following four different models:

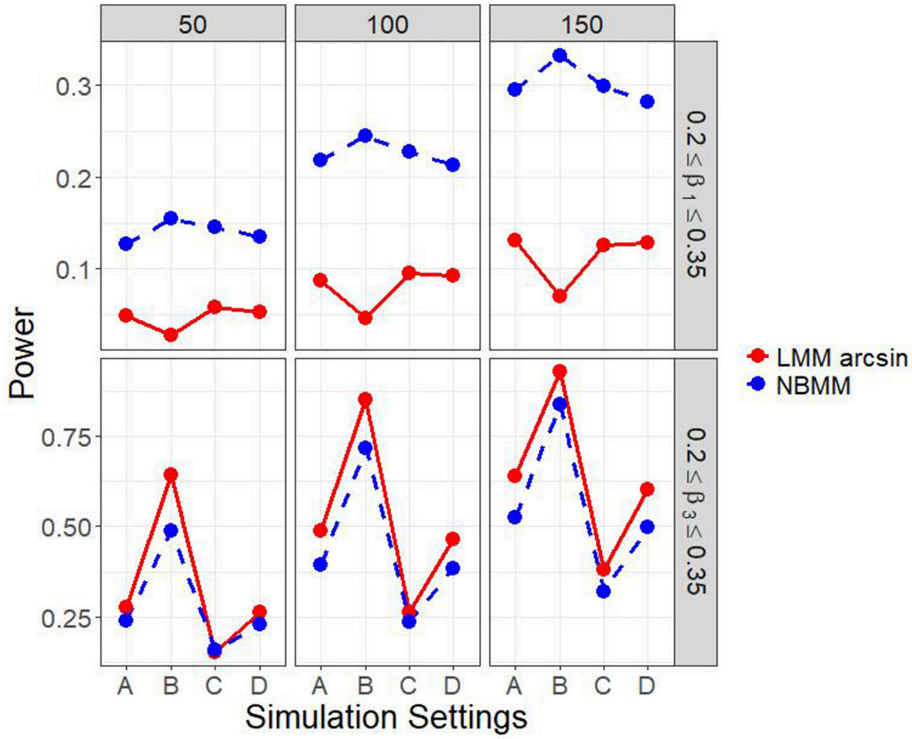


FIGURE 2 | Empirical power of both interaction term and main effect in all four simulation settings.

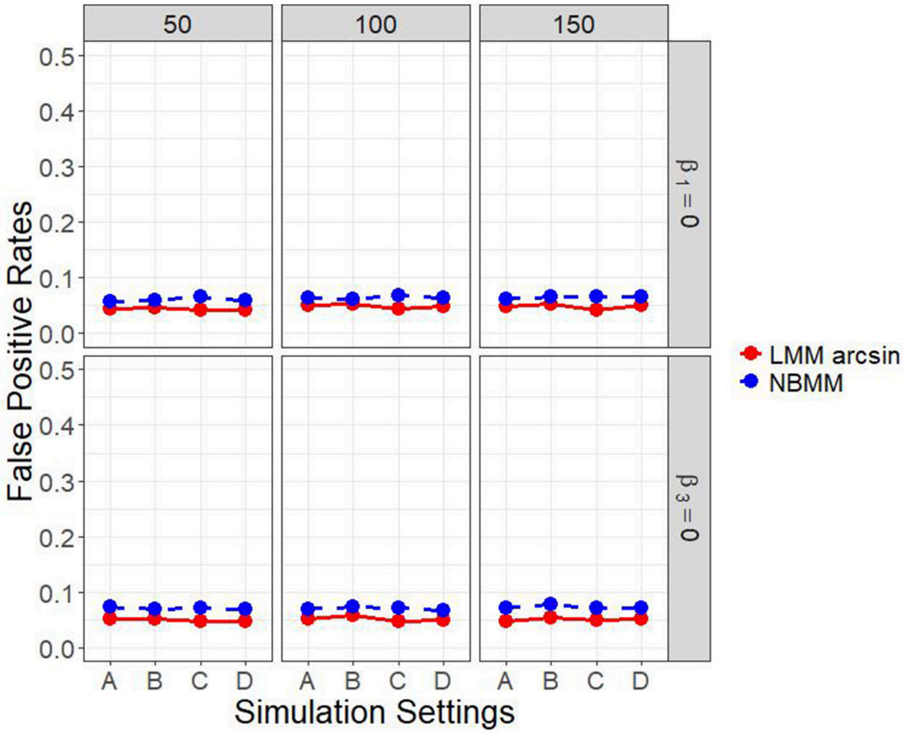


FIGURE 3 | False positive rates of both interaction term and main effect in all four simulation settings.

TABLE 3 | Significant taxa rates detected in four models with LMMs and NBMMs.

		Alpha Level	0.05
Model 1	Test of β_1	LMMs	0.034483
		NBMMs	0.068966
Model 2	Test of β_1	LMMs	0.034483
		NBMMs	0.12069
Model 3	Test of β_1	LMMs	0.12069
		NBMMs	0.224138
	Test of β_3	LMMs	0.137931
		NBMMs	0.275862
Model 4	Test of β_1	LMMs	0.137931
		NBMMs	0.206897
	Test of β_3	LMMs	0.137931
		NBMMs	0.293103

1) Model A: the host factor as fixed effect only, no host factor and time interaction term, only random intercept;
2) Model B: the host factor as fixed effect only, no host factor and time interaction term, two random effects (i.e., random intercept and time effect);
3) Model C: the host factor, time, host factor and time interaction term as fixed effects, only random intercept;
4) Model D: the host factor, time, host factor and time interaction term as fixed effects, two random effects (i.e., random intercept and time effect);

We summarized the number of significant taxa and calculated the rate of significant taxa detected by LMMs and NBMM each using Model A-D at alpha level at 0.05 (**Table 3).** In model A and model B, the numbers of detected significant taxa were substantially less than the numbers from model C and model D. It inferred that failing to incorporate the host factor and time interaction term as fixed effect in the model will largely affect our ability to detect shifts in microbiome studies. Meanwhile, it showed that our NBMMs is capable in detecting more significant taxa than LMMs. Consistent differences have also been found at different significance levels, like 0.01 and 0.001.

Figure 4 shows the significant features of species level in the model with the host factor and the host factor and time interaction term both at the 5% significance threshold and their minus log transformed p-values for NBMMs and LMMs. It showed that NBMMs could discover more species than LMMs in detecting both static association (with host factor term) and dynamic association (with host factor and time interaction term). To compare our analysis results with the published results in DiGiulio et al. (2015), we found that the original paper made two extreme assumptions to the longitudinal study as completely independent or averaged over samples for each subject. The top identified taxa overlapped between our NBMMs with the original paper included *Gardnerella_137183, Lactobacillus jensenii_31171, Staphylococcus aureus_4446058, Lactobacillus crispatus_4447432, Prevotella_760967, Dialister_1105876*. In summary, our NBMMs method is not only a statistical valid method without making extreme assumptions and data transformation, but also detected more significant taxa and yielded much smaller p-values than the LMMs, showing that the proposed method could be more powerful than the conventional LMMs.

DISCUSSION

The main research interest in longitudinal microbiome study is to detect the associations between host clinical/environmental factors and the dynamic shifts in microbiome composition while accounting for sources of heterogeneity and dependence in microbiome measurements. To study the dynamic composition of microbiome, many studies collect samples with temporal structures (Hill et al., 2010; Morrow et al., 2013; Srinivas et al., 2013; La Rosa et al., 2014; Leamy et al., 2014; Faust et al., 2015; Wang et al., 2015; Zhou et al., 2015). These longitudinal studies enable us to study the inherent dynamic properties in microbiome data which have provided extraordinary opportunities to elucidate the true roles of the microbiome in health and disease states and to develop new diagnostics and therapeutic targets (Knights et al., 2011; Segata et al., 2011; Virgin and Todd, 2011; Collison et al., 2012). Accurately identifying and understanding these associations is critical to further predict the probabilities of disease with the identified taxa or biomarkers. However, the traditional methods of using LMMs to model longitudinal data fail to address the count data features in microbiome data. Our simulation studies revealed the impact of the specific features on the microbiome data, showing that ignoring those features can substantially reduce the power for detecting the effects of host clinical/environmental factors with dynamic effects, thus leading to biased and false inferences. We extended our previously proposed negative binomial mixed model (NBMMs) specifically to directly analyze longitudinal microbiome count data without data transformation.

The previously proposed NBMMs (Zhang et al., 2017) have demonstrated its superior ability in family structured clustered microbiome count data. The proposed NBMMs directly model microbiome counts generated by the 16S rRNA gene sequencing or the shotgun sequencing with an efficient IWLS algorithm (Schall, 1991; Breslow and Clayton, 1993; McCulloch and Searle, 2001; Venables and Ripley, 2002). It not only addresses statistical challenges of over-dispersion and varied total reads in microbiome count data, but also accounts for correlation among the observations. Our simulations and real data analysis also show that our algorithm is stable and efficient (Zhang et al., 2017). Meanwhile, the IWLS algorithm is an extension of a commonly used procedure for fitting GLMs and GLMMs which allows us to model non-constant variances or special correlation structures. Therefore, by extending the NBMMs to analyze longitudinal microbiome count data, we illustrated the capability of our proposed NBMMs to handle complex longitudinal study design, such as to include time in the random slope model or to account for the auto-regressive residual correlation in time-series data.

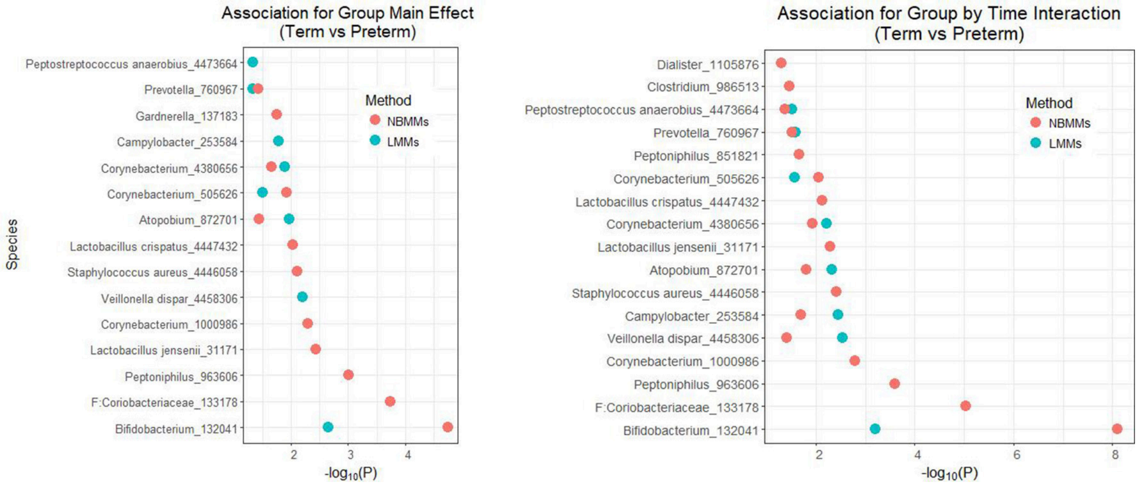


FIGURE 4 | The analyses of NBMMs and LMM: minus log transformed *p*-values for the significant differentially abundant taxa at the 5% significance threshold between term and preterm groups for species level; the left panel shows the minus log transformed *p*-values for association of group main effect, and the right panel shows the minus log transformed *p*-values for association for group by time interaction.

Our simulations indicate that our proposed approach is flexible to handle complex structured longitudinal data, allowing for incorporating any types of random effects and within-subject correlation structures (Pinheiro and Bates, 2000; McCulloch and Searle, 2001). In the simulations, our proposed approach outperformed LMMs consistently.

We also applied our method to a previously published data set. The purpose of the real data is to detect host factors that associated with dynamic compositional features of the microbiome (Leamy et al., 2014). Notably, by applying our NBMMs to the temporal and spatial dataset from DiGiulio et al. (2015), the goal of our analysis was to detect taxa that are significantly associated with dynamic change in compositional microbiome between termed and preterm pregnancy. Our proposed method detected the same species *Gardnerella_137183, Lactobacillus jensenii_31171, Staphylococcus aureus_4446058, Lactobacillus crispatus_4447432, Prevotella_760967, Dialister_1105876,* as in the original paper. In the original paper, they made two extreme assumptions to the longitudinal study as completely independent or averaged over samples for each subject. Our NBMMs, on the other hand, does not make any extreme assumption and is more statistically valid. Nevertheless, we still identified overlapped species as in the original paper, showing NBMMs picked out the significant species under extremes as well. Our NBMMs method detected more significant taxa and yielded much smaller *p*-values than the LMMs, showing that the proposed method could be more powerful than the conventional LMMs. Furthermore, comparing the species identified in the real data using LMMs and NBMMs, we found that the species identified by NBMMs only are mostly overlapped with the original paper. It inferred that the transformation of count data could potentially lead to misleading information and interpretation. One potential limitation of our NBMMs is that it is not designed to explicitly handle zero-inflation and we recommend it as future work. Even though, our NBMMs has shown it outperformed LMMs in longitudinal microbiome study in terms of power and accurate interpretation. It is also directly applicable to be used as an analytic tool in longitudinal RNA-seq study.

AUTHOR CONTRIBUTIONS

NY design the study, develop the method and the software, and participate in writing the paper; XZ simulation study, real analysis, and draft the manuscript; Y-FP design the study, real data analysis, and participate in writing the paper; LZ design the study, and participate in writing the paper; BG design the simulation and real data analysis; AP participate in revising the manuscript; WZ real data analysis, and participate in revising the manuscript.

ACKNOWLEDGMENTS

This work was supported in part by research grants from USA National Institutes of Health (R03-DE024198, R03-DE025646), National Natural Science Foundation of China (31771417), National Natural Science Foundation of China (81673448), and Natural Science Foundation of Jiangsu Province China (BK 20161218).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01683/full#supplementary-material

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

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., et al. (2010). Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* 5:e10667. doi: 10.1371/annotation/df45912f-d15c-44ab-8312-e7ec0607604d

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16:191. doi: 10.1186/s13059-015-0759-1

Breslow, N. E., and Clayton, D. C. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc*. 88, 9–25.

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

Collison, M., Hirt, R. P., Wipat, A., Nakjang, S., Sanseau, P., and Brown, J. R. (2012). Data mining the human gut microbiota for therapeutic targets. *Brief Bioinformatics* 13, 751–768. doi: 10.1093/bib/bbs002

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

De Muinck, E. J., Lagesen, K., Afset, J. E., Didelot, X., Ronningen, K. S., Rudi, K., et al. (2013). Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics* 14:81. doi: 10.1186/1471-2164-14-81

DiGiulio, D. B., Callahan, B. J., Mcmurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11060–11065. doi: 10.1073/pnas.1502875112

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11971–11975. doi: 10.1073/pnas.1002601107

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

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

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

Gerber, G. K. (2014b). "Longitudinal Microbiome Data Analysis," in *Metagenomics for Microbiology*, eds J. Izardm (Cambridge, MA: Academic Press).

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* 12:271. doi: 10.1186/1471-2105-12-271

Gilbert, J. A., Meyer, F., and Bailey, M. J. (2011). The future of microbial metagenomics (or is ignorance bliss?). *ISME J.* 5, 777–779. doi: 10.1038/ismej.2010.178

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., et al. (2016a). Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* 19, 731–743. doi: 10.1016/j.chom.2016.04.017

Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G., and Ley, R. E. (2016b). Cross-species comparisons of host genetic associations with the microbiome. *Science* 352, 532–535. doi: 10.1126/science.aad9379

Hill, D. A., Hoffmann, C., Abt, M. C., Du, Y., Kobuley, D., Kirn, T. J., et al. (2010). Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunol.* 3, 148–158. doi: 10.1038/mi.2009.132

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., and Nicholson, J. K. (2011). Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol.* 19, 349–359. doi: 10.1016/j.tim.2011.05.006

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

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108, (Suppl. 1), 4578–4585. doi: 10.1073/pnas.1000081107

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

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12522–12527. doi: 10.1073/pnas.1409497111

Leamy, L. J., Kelly, S. A., Nietfeldt, J., Legge, R. M., Ma, F., Hua, K., et al. (2014). Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice. *Genome Biol.* 15:552. doi: 10.1186/s13059-014-0552-6

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538. doi: 10.1186/1471-2105-11-538

McCulloch, C. E., and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Hoboken, NJ: John Wiley & Sons, Inc.

Morrow, A. L., Lagomarcino, A. J., Schibler, K. R., Taft, D. H., Yu, Z., Wang, B., et al. (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* 1:13. doi: 10.1186/2049-2618-1-13

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., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Peng, X., Li, G., and Liu, Z. (2015). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* 23, 102–110 doi: 10.1089/cmb.2015.0157

Pinheiro, J. C., and Bates, D. C. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer Verlag.

Pookhao, N., Sohn, M. B., Li, Q., Jenkins, I., Du, R., Jiang, H., et al. (2015). A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes. *Bioinformatics* 31, 158–165. doi: 10.1093/bioinformatics/btu635

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Samuel, B. S., and Gordon, J. I. (2006). A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10011–10016. doi: 10.1073/pnas.0602187103

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* 78, 719–727. doi: 10.1093/biomet/78.4.719

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

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

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* 9, 279–290. doi: 10.1038/nrmicro2540

Srinivas, G., Moller, S., Wang, J., Kunzel, S., Zillikens, D., Baines, J. F., et al. (2013). Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.* 4:2462. doi: 10.1038/ncomms3462

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

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer Verlag.

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

Wagner, B. D., Robertson, C. E., and Harris, J. K. (2011). Application of two-part statistics for comparison of sequence variant counts. *PLoS* ONE 6:e20296. doi: 10.1371/journal.pone.0020296

Wang, J., Kalyan, S., Steck, N., Turner, L. M., Harr, B., Kunzel, S., et al. (2015). Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. *Nat. Commun.* 6:6440. doi: 10.1038/ncomms7440

Ward, D. V., Scholz, M., Zolfo, M., Taft, D. H., Schibler, K. R., Tett, A., et al. (2016). Metagenomic sequencing with strain-level resolution implicates uropathogenic *E.* coli in necrotizing enterocolitis and mortality in preterm infants. *Cell Rep.* 14, 2912–2924. doi: 10.1016/j.celrep.2016.03.015

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

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

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

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

Zhou, Y., Shan, G., Sodergren, E., Weinstock, G., Walker, W. A., and Gregory, K. E. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLoS ONE* 10:e0118632. doi: 10.1371/journal.pone.0118632

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 Microbiology

REVIEW
published: 30 January 2018
doi: 10.3389/fmicb.2018.00035

Network Analyses in Plant Pathogens

David Botero[1,2,3], Camilo Alvarado[1], Adriana Bernal[4], Giovanna Danies[5] and Silvia Restrepo[1*]

[1] Laboratory of Mycology and Plant Pathology (LAMFU), Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia, [2] Grupo de Diseño de Productos y Procesos, Department of Chemical Engineering, Universidad de Los Andes, Bogotá, Colombia, [3] Grupo de Biología Computacional y Ecología Microbiana, Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia, [4] Laboratory of Molecular Interactions of Agricultural Microbes, LIMMA, Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia, [5] Department of Design, Universidad de Los Andes, Bogotá, Colombia

Even in the age of big data in Biology, studying the connections between the biological processes and the molecular mechanisms behind them is a challenging task. Systems biology arose as a transversal discipline between biology, chemistry, computer science, mathematics, and physics to facilitate the elucidation of such connections. A scenario, where the application of systems biology constitutes a very powerful tool, is the study of interactions between hosts and pathogens using network approaches. Interactions between pathogenic bacteria and their hosts, both in agricultural and human health contexts are of great interest to researchers worldwide. Large amounts of data have been generated in the last few years within this area of research. However, studies have been relatively limited to simple interactions. This has left great amounts of data that remain to be utilized. Here, we review the main techniques in network analysis and their complementary experimental assays used to investigate bacterial-plant interactions. Other host-pathogen interactions are presented in those cases where few or no examples of plant pathogens exist. Furthermore, we present key results that have been obtained with these techniques and how these can help in the design of new strategies to control bacterial pathogens. The review comprises metabolic simulation, protein-protein interactions, regulatory control of gene expression, host-pathogen modeling, and genome evolution in bacteria. The aim of this review is to offer scientists working on plant-pathogen interactions basic concepts around network biology, as well as an array of techniques that will be useful for a better and more complete interpretation of their data.

Keywords: networks, bacterial pathogens, plant pathogens, host-pathogen interactions, pathogenicity

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Shiri Freilich,
Tel Aviv University, Israel
Naresh Singhal,
University of Auckland, New Zealand

****Correspondence:***
Silvia Restrepo
srestrep@uniandes.edu.co

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

Received: *22 September 2017*
Accepted: *09 January 2018*
Published: *30 January 2018*

Citation:
Botero D, Alvarado C, Bernal A, Danies G and Restrepo S (2018) Network Analyses in Plant Pathogens. Front. Microbiol. 9:35. doi: 10.3389/fmicb.2018.00035

INTRODUCTION

Biology has entered a new era of scientific discoveries as a consequence of the development of new technologies, and the production of massive amounts of biological data at the cellular and subcellular levels. Researchers can now formulate new hypotheses and diverse manners of testing them. They can design new experiments based on multiple environmental, temporal, and physiological conditions on a single cell, populations, or communities of species. The reduction in costs of next-generation sequencing (NGS) technologies coupled with the advances in metabolomics and proteomics has made high-throughput data more accessible (Hou et al., 2015).

The levels of information that can be obtained from biological entities range from genes, genomes, transcriptomes, proteomes, and metabolomes to phenotypes.

Despite these advances, the amount of collected data is larger than the amount being analyzed. Molecular biologists tend to focus on a single level of information (e.g., specific genes, protein-protein interaction, etc.), ignoring the different levels of interactions and connections present within complex biological systems. In the case of quantitative experiments in the areas of genomics and transcriptomics, the amount of available data exceeds the capacity of the common computational systems as well as the ability for researchers to interpret them. Thus, the challenge resides in building models that accurately represent nature and gaining biological insights from data that is inherently noisy and heterogeneous. Systems biology attempts to bridge this multi-level understanding of living systems (Karr et al., 2012).

Systems biology is a discipline that studies biological entities as a whole. Here, parts of the organism (genes and their regulation, signaling cascades, interacting proteins, structural compounds, and metabolic pathways) interact among them and with the environment (which, in turn, gives a context to the organism) to produce a given phenotype. When the biological parts of an organism are interconnected, new properties arise that are dependent on the context and the biological system. Systems biology uses different sources of biological data, mathematical approaches, and computational methods and techniques, to model the organism in a computer (*in silico*). The computational model allows researchers to make predictions and to generate new hypotheses that may then be experimentally validated. Experimentation can fulfill this function and serve to better parameterize and tune different theoretical models.

One of the fields within systems biology which has been fundamental for studying biological organisms at a large-scale is network analysis. Network analyses are a set of mathematical and computational approaches that may be used to study the interactions between the components of a network such as computers connected through the internet, electrical nodes within a network, or biological components within an organism. In the context of biology, the network approach or network biology allows to reconstruct molecular interactions and uncover biological properties that may be difficult to uncover when studying a single or few interactions.

This review presents the main approaches in network biology and their complementary experimental assays used to investigate bacteria-plant interactions. When necessary, examples of human-pathogen interactions were included to illustrate analyses that may potentially be applied to study plant-pathogen interactions. Pathogenicity is an ecological interaction influenced by many different factors. Understanding molecular and ecological interactions may help explain the mechanisms by which pathogens colonize their host plant as well as the co-evolutionary history among the two or more-interacting species.

The review is divided into five sections. First, we describe the basic concepts in network biology; second, we illustrate the importance of metabolic pathways in bacterial pathogenicity; third, we review different approaches used to study protein-protein interactions; fourth, we review the modeling of regulatory networks; and fifth, we describe how this information, may be used to understand processes of adaptation of pathogens to recent and former hosts. The aim of this review is to offer scientists in the field of host-pathogen interactions, the most important concepts around network biology, as well as an array of techniques that will be useful for a better and more complete interpretation of their data.

NETWORK ANALYSES IN SYSTEMS BIOLOGY

Network biology has arisen as a new subfield of systems biology (**Box 1**) useful in molecular biology studies. The high amounts of data produced by omics technologies nowadays, as well as the increasing number of studies on bacterial pathogenesis allows the use of network biology to mathematically model large-scale bacterial systems. Network biology, is a top-down approach (**Box 1**) that allows the reconstruction of genome-scale biological systems.

The biological networks represent the relationships among molecular components within the context of a cellular function (**Box 2**). The methods derived from the mathematical framework of networks can be applied to diverse fields such as electrical, social, and Internet networks. As biological systems can be represented as networks, the mathematical concepts behind network analyses can be applied to biomolecular systems.

Types of Biological Networks

In the context of networks and molecular biology, we can represent an organism, or parts of it, using four different kinds of networks: regulatory, metabolic, protein-protein interaction networks (PPINs), and signaling networks (a special type of PPIN). Furthermore, these networks can be integrated into a single model by using a combination of different networks connected into a single computational model. It is important to note that the classification of regulatory, metabolic, and PPINs is arbitrary and has been done to facilitate the construction of scientific knowledge. This review, focuses on these three methods due to the availability of omics data from pathogens that may be used to construct these types of networks. The omics data that have been generated have been mostly used to investigate specific research questions, leaving large amounts of data yet to be explored. Network analyses provide an opportunity to further analyze this information to develop new hypotheses related to mechanisms of pathogenesis or general life style of these microorganisms.

One type of network is the transcriptional regulatory network (TRN). TRNs are used to mathematically represent gene expression profiles and their regulation by transcription factors or other regulatory elements (e.g., sRNA). Through these TRNs, one can simulate the effect of different biological and environmental conditions on the expression profile of an individual. The TRNs may be constructed for specific groups of genes, such as those related to pathogenicity, or for the whole organism. In a topological sense, the TRN is defined as a bipartite network (**Box 2**) with directionality. Some nodes correspond

BOX 1 | Systems biology.

Systems biology comprises different combinations of mathematical and computational approaches used with diverse kinds of the biological data; as a result, the starting scale of the model (whether it takes into account a small subsystem or a whole system) will vary. Therefore, systems biology may use two approaches that are complementary depending on the nature of the data, and the mathematical and computational approaches used: **bottom-up** and **top-down** approaches (Bruggeman and Westerhoff, 2007).

The **bottom-up** approach precisely reconstructs biological subsystems from their parts (genes, proteins, and metabolites) until a full model of the subsystem is obtained (mainly at a small scale). This kind of approach allows to **deduce** fundamental principles inherent to all biological systems such as the physical and mathematical laws that govern it. The data used for the model are obtained from single cell experiments, from the *in vitro* assessment of rate parameters from enzymatic reactions, transport phenomena, or regulatory processes.

The **top-down** approach reconstructs the biological system from high amounts of data to initially obtain a full draft model of the whole system, with subsequent refinings. This kind of approach allows to **induce** properties of the system in a biological state. The data used for these models arise from omics experiments (genomics, transcriptomics, metabolomics, etc.), and they allow the reconstruction of the whole model.

BOX 2 | Biological networks.

A biological network is a mathematical abstraction of nature which represents biological entities such as genes, transcription factors, metabolites, and proteins, as **nodes** or **vertices** and the relations between them as **edges** or **links** (regulatory mechanism, transformation reactions, protein-protein interactions such as signaling cascades). We can find **directed** (regulation of A to B, directionality of an enzymatic reaction, **Figure 1A**) or **undirected** (a pair of interacting proteins, **Figure 1B**) networks, depending on whether the relationship between the nodes has directionality or not, or if this can be determined. Furthermore, there are **unipartite** networks where the nodes have the same biological feature (e.g., protein-protein interaction networks) and **bipartite or two-mode** networks, composed of different biological components (e.g., a regulatory network where regulatory proteins and regulated genes interact or, metabolic networks where substrates are connected to reactions and reactions with substrates) (Newman, 2010).

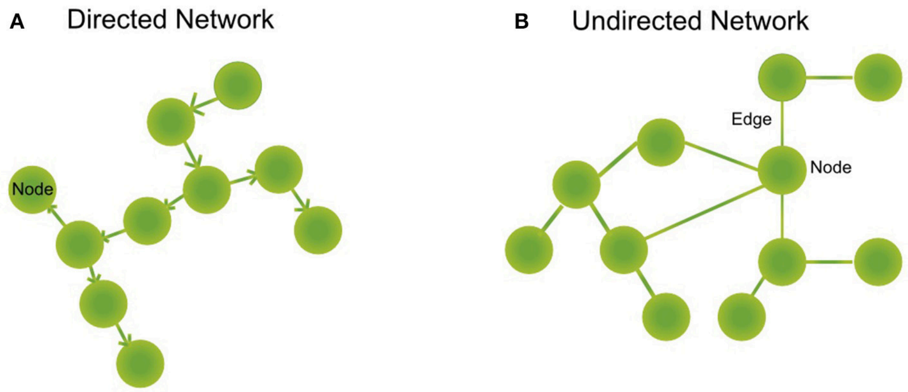


FIGURE 1 | Type of networks. **(A)** Directed networks are composed of nodes representing biological entities as proteins, metabolites, or genes. These nodes are interconnected by directed edges (or arrows) that symbolize a directed relationship between two or more biological species, as a gene regulated by a transcription factor or a reaction that is connected downstream to another reaction forming metabolic pathways. **(B)** Undirected networks are composed of nodes, that represent proteins, for example. These nodes are interconnected by edges that symbolize an interaction between two or more biological species, as for example signaling proteins.

to regulatory proteins and others to target genes (that can be transcriptionally switched on or off by the regulatory protein). One regulatory protein can be connected to several target genes; in turn, genes can be regulated and connected by one or a small number of regulatory proteins.

Metabolic networks are substrate-product transformation networks mediated by enzymatic reactions. In the metabolic networks, the substrates and products can be proteins, lipids, and other cellular components. These are represented as nodes and the transformation reactions mediated by enzymes are represented as edges. This representation of metabolic networks can be analyzed by computational methods to perform associations between the genotype and the metabolic phenotype of an organism, as constraint based modeling does (**Box 3**). Metabolic networks may be coupled to the regulatory networks of an organism to model a more complex representation of the molecular machinery of the organism.

A PPIN reflects physical interactions between two or more proteins. In this category, we can find signaling networks, but it is also possible to find proteins involved in the formation of macromolecular complexes related to structural and molecular types of machinery of the cell. The Signaling Network contains a series of proteins that are transformed to carry a signal inside or outside of the cell. Signaling cascades are of special interest

BOX 3 | Systems biology.

The metabolism of an organism may be represented in a matrix based on the stoichiometry of the reactions in the **constraint-based modeling** (CBM) approach (Orth et al., 2010). The stoichiometric matrix can be analyzed to assess the metabolic phenotype of the organism under different conditions (e.g., environment, mutants, etc.). To analyze the metabolic phenotype, the stoichiometric matrix may be solved using a Flux Balance Analysis (FBA). A FBA is a computational optimization method. The final solution of the metabolic system is the distribution of the reaction rates or fluxes (moles over time). In the FBA, assumptions and constraints of the system are defined. For example, it assumes a steady-state (thermodynamic equilibrium) and defines upper and lower boundary constraints for the fluxes throughout the reactions. Furthermore, an objective function must be defined to achieve a unique solution of the system. The **objective function** is a reaction or a combination of reactions that represent a biological feature of the organism e.g., biomass. In other words, models based on CBM approach represents the metabolism of an organism only with the information of the reactions catalyzed by enzymes that are coded in the genome.

Another alternative approach that does not require the calculation of an optimal flux distribution is the **elementary flux mode analysis** (EFMA) (Zanghellini et al., 2013). In this analysis, the metabolic network is decomposed in its main pathway components.

A complementary analysis in metabolic modeling is **gene set enrichmen**t **analysis** (GSEA) (Hung et al., 2012). When applied to a genome-scale, set of genes differentially expressed can be classified into metabolic categories or pathways giving information related with the most represented pathways in a determined scenario.

in molecular pathosystems since they are tightly related to the regulation of the response to attack and defense of the pathogen and the host, respectively.

When a biological network follows the power law distribution several biological interpretations based on the network metrics can be stated (**Box 4**). However, these interpretations must be carefully reviewed from the biological point of view of the researcher. We recommend the work of Winterbach et al. (2013), which provides detailed description of these statistics (Winterbach et al., 2013). For a more extensive revision of the mathematical foundations of biological networks, please refer to De Smet and Marchal (2010), Képès (2007), and Newman (2010).

METABOLIC NETWORKS AND PATHOGENICITY

In this section, we will review studies on metabolic modeling of plant pathogenic bacteria. Given that the information may be limited, we will also include examples of animal pathogens. First, we will describe the constraint-based modeling (CBM) approach, commonly used for *in silico* metabolic modeling. Second, we will review the biological results produced by these studies and their main conclusions; of special interest will be the objective function. Third, we will review the multiscale metabolic modeling approach that integrates different sources of data and constraint-based metabolic models. Finally, we will discuss how CBM is a hypothesis-driven approach used in metabolic networks and the possibilities to improve metabolic models based on experimental results.

Constraint-Based Modeling

The metabolic interactions within an organism can be modeled and analyzed using different mathematical approaches, among others, deterministic kinetic models, stochastic models, elementary flux mode analysis, CBM, and pathway enrichment analysis (**Box 4**; Puchałka and Kierzek, 2004; Hung et al., 2012; Zanghellini et al., 2013). Every one of these methods has advantages and disadvantages. The CBM approach has been established as a standard for metabolic model formulations; there are approximately 165 models of organisms that are finished and experimentally validated (http://sbrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms). This method has been widely employed given that it is a top-down approach (**Box 1**) that may incorporate whole-genome data and chemical information that is publicly available, as well as knowledge obtained through experimentation into a genome-scale metabolic model. With this approach, several analyses can be performed and relevant biological questions can be addressed (Oberhardt et al., 2009). Other mathematical approaches, such as the mass-action kinetic model (Horn and Jackson, 1972) or the biochemical system analysis (Savageau, 1969), rely on several parameters such as rates of transformation of molecules involved in metabolic reactions of the cell. These parameters are difficult to calculate experimentally at a whole-genome level (that is, all the possible reactions catalyzed by all the enzymes coded by the genome). Thus, the CBM offers a powerful approach to assess metabolic phenotypes in distinct environmental conditions by relying on physicochemical constraints that restrict the metabolic phenotype[1] of the organism.

The first part of the CBM approach is the genome-scale reconstruction of the metabolic network (**Figure 2**). There are five main steps in obtaining a high-quality metabolic reconstruction of an organism: (i) genome annotation; (ii) gene-protein and protein-reaction associations; (iii) model curation; (iv) validation through experimental analyses; and (V) improvement of the metabolic model by incorporating the feedback obtained through experimentation.

Genome annotation can be performed using different bioinformatics tools, such as the Rapid Annotation using System Technology (RAST; Aziz et al., 2008; Richardson and Watson, 2013; Kalkatawi et al., 2015). After the genomes have been automatically annotated, they must be manually curated. Once a high-quality genome annotation is obtained, proteins involved in metabolic reactions are assigned. Commonly used databases for the assignment of proteins to metabolic pathways include the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), MetaCyc (Caspi et al., 2014), MetaNetx (Ganter et al., 2013), and Biochemical, Genetic and Genomic (BiGG) knowledge base (Schellenberger et al., 2010).

[1] The metabolic phenotype is the distribution of the biochemical reactions rates (fluxes) in a determined set of conditions (physicochemical constraints).

BOX 4 | Topological features of biological networks.

The most basic topological measure of a network is the **average (or mean) degree**, that is the average number of nearest nodes connected to a specific node. More informative is the distribution of these nearest nodes in a network. A feature of some types of biological networks, particularly in metabolic and protein-protein interaction networks is that they follow a **power law** distribution (or a **scale-free** degree distribution). That is, most of the nodes have few neighbors and the minority of them have hundreds or thousands of neighboring nodes (highly connected nodes or **hubs**) (Barabási and Albert, 1999). Particularly the biological networks have a specific topology that tends to be small-world; in other words, their organization makes possible the existence of hubs as central focal points of interactions (Aloy and Russell, 2004), e.g., proteins or genes involved in regulatory processes of bacterial pathogenicity or plant resistance (Haynes et al., 2006).

Another powerful structural measure is the **clustering coefficient**, which evaluates the degree of grouping between a node and its neighbors. The clustering coefficient is defined as the ratio of the number of connections between the neighbors of a node and the total number of possible connections (Watts and Strogatz, 1998). This coefficient measures the **modularity** of a set of nodes and interestingly this feature has shown some patterns of hierarchy between clusters of nodes (modules) of known metabolic networks (Ravasz et al., 2002). Finally, the motifs are patterns of connections between a few nodes that are related to a biological function, particularly in regulatory networks (Maslov and Sneppen, 2002). In topological terms, motifs are sets of nodes whose pattern is overrepresented in the network when compared to a randomly generated network of the same size. In **Table 1** we describe other relevant measurements and concepts of networks.

Once the reactions related to the organism of interest are obtained, a mathematical representation of the connected reactions (metabolic models or pathways) can be reconstructed (Orth et al., 2010). However, this initial representation is not free of gaps and errors. These may arise for different reasons such as an inherent gap in our knowledge of bacterial metabolism (e.g., protein-reaction associations), the incomplete genome sequencing of the organisms, or the inaccuracy in the genome annotation. Therefore, the metabolic model needs to be subjected to a curation process. Several methods and algorithms have been developed to curate this model (e.g., based on homology and phylogenetic information or experimental data) (Orth and Palsson, 2010).

Some useful automatic tools that can be alternatively used to reconstruct and analyze metabolic networks include RAST-SEED (Aziz et al., 2008), KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007), Reconstruction, Analysis, and Visualization of *Metabolic* Networks (RAVEN) (Agren et al., 2013), PRofils pour l'Identification Automatique du Métabolisme (PRIAM) (Claudel-Renard et al., 2003), SuBliMinal, and Pathway Tools (Swainston et al., 2011). Furthermore, protocols for supervised and manual reconstruction of metabolic networks have been established (Francke et al., 2005; Reed et al., 2006; Thiele and Palsson, 2010; Pinzón et al., 2011; Lewis et al., 2012).

Once a representation of the metabolism of the bacterium is obtained, relevant biological questions can be addressed based on this model. For example, the rate of ATP production or the oxygen consumption can be assessed. In the CBM approach, several constraints are set to assess the metabolic model of the organism (McCloskey et al., 2013). The metabolic phenotype can be defined as the rates of consumption and production of the metabolites for every reaction of the metabolic model of interest in a determined biological context or environment.

Constraints are determined *a priori* based on either experimental or theoretical data like metabolomics, C^{13} labeling and measurements of consumption and production of carbon sources. An example may be the active and inactive reactions that reflect the biological state of the cell and can be determined, although indirectly, through specific transcriptional profiles (genes down or up-regulated). Another example of constraint includes the activation of transport reactions that simulates the substrate transported into the cell in a specific medium or biological condition. Therefore, the metabolic phenotype, which is defined by a set of reactions that represents a biological function, such as growth or pathogenicity, can be assessed.

Flux Balance Analysis (FBA) is an approach used in CBM to find an optimal distribution of the rates of conversion of substrates to products (fluxes), in every reaction. In order to obtain the solutions for the reaction rates of interest, a representation of a specific biological function must be defined (e.g., growth, redox potential, production of a compound of biological, or industrial interest, etc.). This specific biological function is known as the objective function. Choosing the best objective function to answer a specific biological question is still controversial. The right choice will define the robustness of the conclusions achieved by the computational analysis (see discussion below). Finally, FBA allows uncovering the most reliable mechanism behind a relevant biological function (O'Brien et al., 2013).

Metabolic Modeling of Pathogenic Bacteria

As mentioned above, the constraint-based modeling, CBM, has been established as a standard method for modeling the metabolism of microorganisms (especially in bacterial pathogens), given that it only relies on a few physicochemical constraints and on the assumption that the metabolic fluxes of the organism are in a steady-state (**Box 3**). With this approach, metabolic phenotypes of pathogenic bacteria may be simulated. Such simulations may reflect differences between wild-type bacteria and their mutant derivatives, between pathogenic and non-pathogenic bacteria, and the effect of growth at different environmental conditions.

The main biological questions addressed in metabolic models of plant pathogens, using CBM, are related to the search for control strategies against these pathogens, the classification of pathogens, the comparisons between pathogenic and non-pathogenic strains, and the plant-pathogen interactions. The CBM approach allows studying the metabolism of pathogens for the search of alternative strategies for control, and through several *in silico* and experimental approaches, has aimed to reveal the metabolic mechanisms, genes, and proteins that are important for pathogenicity. An example is the study of xanthan, a virulence factor of industrial importance, in *Xanthomonas*

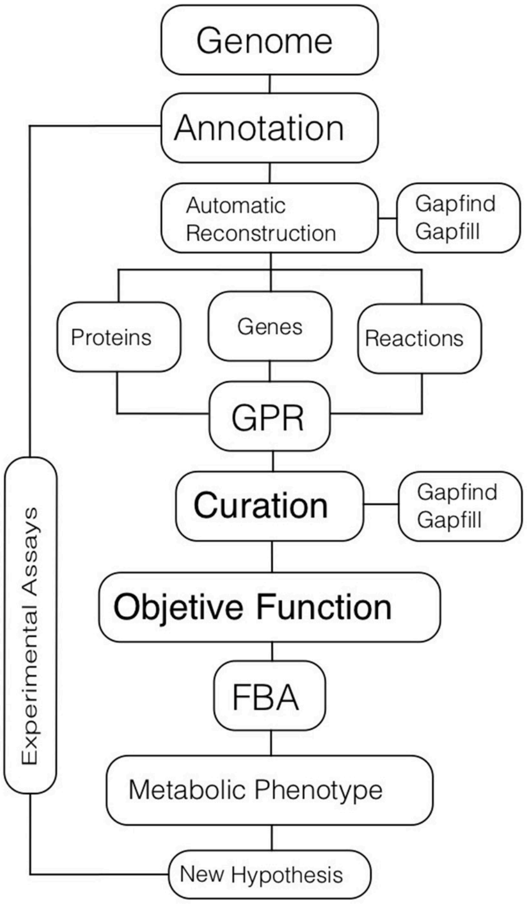


FIGURE 2 | Metabolic modeling. The process of metabolic modeling starts with a genome annotation used for inferring metabolic reactions that are present in an organism. Automatic tools could be used for reconstructing the metabolic network based on the genome. In the initial set of reactions there will be metabolic gaps or missing reactions that are necessary for the complete function of pathways. These gaps can be identified and filled out using different algorithms. The final metabolic reconstruction will have associations among genes, proteins, and reactions (GPRs). Then, further manual curation, based on omics data and literature should be performed. The definition of an objective function that represents a target biological function to optimize should be defined, typically cell growth or ATP production. Once the objective function is set, computational simulations for obtaining metabolic phenotypes related to different conditions are carried out; Flux Balance Analysis (FBA) is the main technique for these simulations. Finally, new biological hypotheses are generated and validated. In all the procedure, data, and information from different experimental assays are incorporated into the model.

campestris pv. *campestris* (*Xcc*) (Schatschneider et al., 2013). Another example, includes the study of metabolic precursors of lipopolysaccharides in *Pectobacterium carotovorum* because of their role in antimicrobial resistance (Wang et al., 2014a). These two studies highlight the importance of virulence factors in the relocation of resources for pathogen growth and their potential use as drug targets.

Gene essentiality analyses have been used to find genes that are related to pathogenicity through the systematic deletion of every gene related to metabolism. The *in silico* deletion of genes in the whole reaction network allows the identification of important genes for the survival of the pathogen (Segrè et al., 2002; Shlomi et al., 2005; Kim et al., 2007). In the case of *X. campestris* pv. *campestris,* several essential genes were identified *in silico* (Schatschneider et al., 2013). Furthermore, the researchers performed experimental validation by generating mutants of the carbohydrate metabolism and xanthan production. Interestingly in this study, a subset of these genes, that were initially identified as non-essential, were found to cause a meaningful decrease in the growth rate, after additional *in silico* double mutants were performed. This highlights the importance of double mutants for the determination of essential genes and the reduction of false negative results in pathogenicity assessments.

The CBM approach has helped to compare pathogenic and non-pathogenic bacteria (Perumal et al., 2009; Charusanti et al., 2011; Liao et al., 2011; Monk et al., 2013). Correctly classifying and comparing between pathogenic and non-pathogenic bacteria is important because differences between these may help select the best target for pathogen control. Also, the CBM approach can improve our understanding of the emergence of new pathotypes and their adaptation process to different niches (Monk et al., 2013). Thus, pathogenic mechanisms and infection strategies may be revealed through CBM. However, there are also cases where the metabolic model wrongly predicts the ability of different bacterial mutants or strains to grow on different media. The reasons are metabolic reconstruction artifacts such as incomplete genome information and gaps in our knowledge of the metabolism. However, metabolic network reconciliation methods have been developed to improve the level of prediction of the models (Oberhardt et al., 2011). Ultimately, the inclusion of exact metabolic parameters such as rates of metabolic conversion and rates of volume dilution, achieved through bottom-up approaches, will improve the level of prediction of metabolic models at a genome-scale (Bruggeman and Westerhoff, 2007).

Multiscale Metabolic Modeling

Several studies have focused on integrating different omics information (e.g., RNA-Seq, microarrays, metabolomics, etc.) into the metabolic, protein-protein interaction, and regulatory models of pathogens. Also, the metabolic interactions between hosts and pathogens have been subject of study. This integration has improved the phenotypic predictions, the understanding of the mechanisms of host-pathogen interactions, and have helped discover new drug targets in pathogens (Colijn et al., 2009; Bordbar et al., 2010; Ward et al., 2010; Lobel et al., 2012; Schaadt et al., 2013).

Control at the Metabolic Phenotype in Bacterial Pathogens

Different approaches have been proposed for integrating regulatory and metabolic models in bacterial pathogens of humans, these have not been reported so far for plant pathogens. For example, the regulatory network and the CBM model of *Mycobacterium tuberculosis* were integrated using a probabilistic approach; this model was used to predict the growth rates of

TABLE 1 | Basic concepts of biological networks.

Structure assessment	Definition	Utility	References
Degree distribution	Distribution of probabilities of degrees in a specific network.	Comparisons, scale-free networks. Clear indicator of the presence of hubs when it is combined with the centrality measurement. Degree provides clues about modules in a network by determining the number of interactions shared between neighboring nodes.	Képès, 2007
Shortest path	The shortest path between two nodes in a biological network.	Connectivity.	Perumal et al., 2009
Average diameter	The minimum number of edges connecting any two nodes over all possible pairs.	Information flow, Small World. Capacity and time of the response of a system, so that in networks with a high centrality, signaling processes are favored.	Képès, 2007
Node clustering coefficient	The ratio of connections to neighboring nodes to the number of all possible connections.	Comparisons, scale-free, hierarchical.	Képès, 2007
Betweenness—centrality	The ratio of the number of k-shortest paths passing through a node and its nearest neighbor links.	Identifies hubs (highly connected nodes in a network), important in pathogenicity and potential target for drugs. Hubs may potentially disconnect the network if they are removed or blocked.	Goh et al., 2001; Perumal et al., 2009
Assortativity	The probability of connection of a node with others of the same degree.	Robustness to node deletion.	Newman, 2010

Summary of structural measurements of the topology of a network and their utility in a biological context.

different mutants and putative drug targets (Chandrasekaran and Price, 2010).

A similar approach was used in *Listeria monocytogenes* to decipher its metabolic requirements and the relationship between metabolism and virulence regulation (Lobel et al., 2012). The researchers found a correlation between the activity of certain gene regulators, under nutrient limiting conditions, and the activation of a global virulence response.

The integration of regulatory models and the metabolic model combined with experimental data is fundamental for adjusting the predictions of the metabolic phenotype. For example, Bartell and collaborators found that inconsistencies between the growth rate of *Burkholderia* in different carbon sources, that were experimentally measured, and the predictions obtained by the simulations of the metabolic model, could be partially explained by the absence of the integration between a regulatory model and a metabolic model (Bartell et al., 2014). In the previously mentioned studies, of *M. tuberculosis* and *L. monocytogenes*, researchers included in the metabolic model data obtained from experimental techniques such as microarrays, mutants, transcription factor, RT-qPCR, and lux reporters. These examples highlight the importance of experimental feedback and validation of the model for improving computational predictions, and the integration of regulatory networks into metabolic models.

A subsequent step after the coupling of the regulatory and metabolic models is the incorporation of signaling networks into the pathogenic bacterial model. For example, in *Pseudomonas aeruginosa* several genes related to quorum sensing (QS), an important process in pathogenesis that regulate the expression of virulence genes, were modeled through a multi-level approach using a Boolean method of the signaling, regulatory and metabolic networks (Schaadt et al., 2013). In this work, the researchers identified the best targets at the signaling and metabolic level to inhibit the production of auto-inducers and thus, disrupt the cellular communication between bacteria at the QS system level.

Another example of a multilevel model is *Mycoplasma genitalium* (Karr et al., 2012). This was the first effort to construct a whole model of a microorganism. In this study, 28 different submodels were used to represent the life cycle of the bacterium at the regulatory, metabolic, and signaling level. To accomplish this task, four mathematical approaches were used: (i) Ordinary differential equations, (ii) Boolean logic, (iii) probabilistic, and (iv) CBM approach.

Finally, the integration of molecular networks can be used to study microbiome interactions in pathogenic and non-pathogenic bacteria. In a study of two bacterial species, *Clostridium difficile* and *Barnesiella intestinihominis* the interaction at the metabolic level was investigated. The researchers found in their *in silico* analysis that the competition between the two bacteria reduces the growth of one of them at the expense of the other; this result was experimentally validated (Steinway et al., 2015).

Host-Pathogen Interactions

The interaction between hosts and pathogens has been widely studied in human pathogens through network biology. The research focus can be either the pathogen or the host, depending on the biological question. For example, two different studies of the interaction between *M. tuberculosis* and humans were conducted, both based on genome-scale metabolic using a CBM approach. In the first one, researchers exposed the pathogen to human macrophages, human sputum, and other *in vitro* conditions, and then integrated transcriptomics data of each condition into the metabolic model of the pathogen (Bonde et al., 2011). The objective, in this case, was to study the metabolic changes in the pathogen caused by the interaction with the host in a similar way as has been done for regulatory-metabolic networks. In this research, a down-regulation of the central metabolism and an up-regulation of the cell wall and virulence factors in the pathogen were found. In the second study, the objective was to investigate the metabolic changes in the human alveolar cells as well as in the pathogen, *M. tuberculosis* (Bordbar et al., 2010). In this study, transcriptomic data was also used to assess the interaction between the host and the pathogen. Here, the two metabolic models of both the host and the pathogen were integrated. As a result, a reduction of the metabolic plasticity of the host when interacting with the pathogen (in a technical sense: they found a reduction of the solution space of fluxes in the metabolism of the host) was found. Also, the gene essentiality analysis was improved by the incorporation of the interaction in the modeling process.

Other software tools can be used to model metabolic interactions between the host and the pathogen as are the E-flux (Colijn et al., 2009) or NetGenerator (Schulze et al., 2015) approaches. The E-Flux tool extends the genome-scale reconstructions and CBM approach, by integrating transcriptomic data into the model. Using this tool, it was possible to measure the impact of 75 drugs and nutrients on the cell wall synthesis and fatty acid biosynthesis on *M. tuberculosis*, identifying several inhibitors; importantly, some of the drugs tested are among the most widely used in the treatment of this disease (Colijn et al., 2009). The NetGenerator tool allows the incorporation of different time points in host-pathogen interactions. This method has been used to infer regulatory changes between *Candida albicans* and dendritic cells of *Mus musculus* at different time points during the interaction (Schulze et al., 2015).

Plant-pathogen interactions may also be studied through network biology. For example, host-pathogen networks have been constructed using microRNA and PPIN between *Arabidopsis thaliana* and *Xanthomonas campestris* pv. *campestris*. This study provided several potential pathways of pathogenesis (Kurubanjerdjit et al., 2012). Furthermore, the change from healthy state to disease in *A. thaliana* when infected with *Pseudomonas syringae* pv. *tomato* has been assessed (Ward et al., 2010), by integrating data from microarrays and metabolomics techniques and analyses such as: Proton Nuclear Magnetic Resonance (^{1}H-NMR), Flow Injection Electrospray Mass Spectrometry (FIE-MS), Gas chromatography-mass spectrometry (GC-MS) and GC-TOF-MS (TOF by "time of flight"). This study found that the metabolism of sugars is modified in the plant to improve the flow of energy into the bacteria. Other modifications were nitrogen mobilization and purine metabolism. On the other hand, the plant showed an unusual metabolic activity of aromatic amino acids and secondary metabolites (toxins) potentially used as a defense mechanism against the pathogen.

Plant-pathogen interactions have also been modeled completely *in silico*. Duan et al. (2013) investigated five host-pathogen metabolic models. They analyze two main points: the impairment of the plant by the pathogen and the divergence between host and pathogens' networks. They calculated the metabolic impairment of the plant by identifying the metabolites from the plant that, when taken by the pathogen, affect the plant's growth (in other words, modifies the value of the objective function after FBA). The researchers found that the impairment of the plant metabolic network is determined by the pathogen and not by the host. For the comparisons between host-pathogen interactions, the authors used a multidimensional scaling (MDS) analysis. The MDS approach allows the comparison among different types of host-pathogen interactions. Using a Jaccard distance to measure the pairs of metabolic networks, authors found that the five metabolic networks of the plants studied are very similar to each other. In contrast, the pathogen networks are much more heterogeneous among them. For example, the metabolic networks of the bacterial pathogens *Xanthomonas oryzae* and *P. syringae* differed from those of the fungal pathogenic species. Additionally, researchers found that histidine is the main target in all host-pathogen interactions, followed by lysine, methionine, and the nucleotide phosphate TTP; and in the specific case of *X. oryzae*, thymidine triphosphate. They also found that the large secondary metabolism of plants is underrepresented by a gap of knowledge. However, authors recognize a bias in their study as they only compared pathogenic interactions. The solution proposed, is to use, in addition to the plant-pathogenic networks, non-pathogenic interactions as a null model to compare and validate the results found *in silico*. However, how can this *in silico* simulations be contrasted with experimental data? Interactions among non-pathogens and their host may be compared to pathogenic interactions at the metabolic level to add experimental information to *in silico* predictions.

Objective Function in Pathogenic Bacteria

The objective function is indispensable for the CBM approach because it specifies the set of metabolites that must be used to optimize the system and resolve the metabolic fluxes of the organism. The most frequently used objective function to model pathogenic bacteria is biomass (**Table 2**) (Charusanti et al., 2011; Liao et al., 2011; Thiele et al., 2011; Fong et al., 2013; Monk et al., 2013; Schatschneider et al., 2013; Wang et al., 2014a).

When the organism under study lacks experimental data for the formulation of the biomass function, data from *Escherichia coli* is used. However, differences in the composition of biomass of the components should be considered to correct for the growth estimation of the model. For example, the biomass composition of *Klebsiella pneumoniae* has a greater proportion of carbohydrates than that of *E. coli* (probably due to differences in the polysaccharide content of its capsule); this factor was included in the model of *K. pneumoniae* and it led to an improvement in growth predictions for this species (Liao et al., 2011). Similarly, in a study performed with *Burkholderia*, it was found that the special fatty acid and lipid composition of this species was dependent on the growth temperature. Thus, this information was taken into account when determining the biomass composition used for the objective function to improve the growth predictions of this pathogenic bacteria (Bartell et al., 2014).

An important modification to the biomass function is the inclusion of the growth associated maintenance (GAM) and

TABLE 2 | Examples of objective functions used and the biological utility of the studies.

Organisms	Biological question—objectives	Objective function	References
Yersinia pestis CO92	Gene targets for antibiotic development. Growth at different carbon sources (used for classification of strains of *Y. pestis*).	Biomass: at two temperatures. Differences in LPS and fatty acid composition at biomass definition.	Charusanti et al., 2011
Salmonella enterica serovar Typhimurium LT2	Metabolic reconstruction, reconciliation of two models.	Biomass	Thiele et al., 2011
Salmonella enterica serovar Typhimurium	Reconciliation of simulations and experimental data; gap filling.	Biomass	Fong et al., 2013
Pseudomonas putida KY2440 & P. aeruginosa PA01	Search for drug targets and comparison of metabolic networks of pathogenic and non-pathogenic bacterium.	NA	Perumal et al., 2009
Burkholderia cenocepacia j2315 & *B. multivorans* ATCC 17616	Differences and similarities in pathogenesis and virulence.	Biomass: special composition of lipids and fatty acid.	Bartell et al., 2014
Pectobacterium carotovorum PC1	Establishes a new strategy for identification of bactericides targets of agriculture importance.	Biomass: *E. coli*	Wang et al., 2014a
Klebsiella pneumoniae	Metabolic model reconstruction and experimental validation of the model.	Biomass	Liao et al., 2011
Xanthomonas campestris pv. *campestris*	Uncover mechanisms of xanthan biosynthesis for industrial purposes and pathogenicity research.	Biomass/ xanthan production	Schatschneider et al., 2013
Xanthomonas oryzae pv. *oryzae* & *Pseudomonas syringae* pv. *tomato*	Research on plant-pathogen interactions.	NA	Duan et al., 2013
Escherichia coli (55 strains) & *Shigella* (8 species)	Determination of limits between strain and species at a metabolic level. Characterization of **pan** and **core** metabolic capabilities. Evaluation of strain-specific auxotrophies.	Biomass	Monk et al., 2013

non-growth associated maintenance (NGAM) energies (Thiele and Palsson, 2010) as was performed in *K. pneumoniae* (Liao et al., 2011). The GAM is a reaction that represents the energy necessary (ATP) for the replication of the cell including DNA, protein, and RNA synthesis. The NGAM represents the energy necessary (also in ATP) for maintenance of the cell in activities other than growth (e.g., turgor pressure or membrane leakage). The objective is to adjust the model to the experimental growth data and to account for the differences among strains (Varma and Palsson, 1994).

Another objective function that has been used for pathogenic bacteria are virulence factors. Xanthan, in *X. campestris* pv. *campestris* (*Xcc*) was chosen with excellent results (Schatschneider et al., 2013). The main difficulty for the model of *Xcc* under the phenotype of xanthan was the lack of information in the metabolic databases regarding the polysaccharide biosynthesis needed for xanthan production. This gap was filled by Schatschneider et al. (2013) using additional information from the genome annotation performed in a former study (Vorhölter et al., 2008). Another problem detected by Schatschneider et al. (2013) was that the biomass function competes for the same precursors as xanthan. Thus, for the analysis, xanthan may be defined as a product along with biomass in a specific ratio. The most important result of this study was the discovery of an increased growth rate in the absence of xanthan production by a reallocation of carbohydrate precursors to the biomass products. Finally, the authors validated this prediction using experimental mutants of the carbohydrate metabolism and xanthan production (Schatschneider et al., 2013).

Bartell et al. (2014) extensively assessed the production of several virulence factors of *Burkholderia* species during cystic fibrosis in humans by *in vitro* and *in silico* assays. The virulence factors included biofilm-related exopolysaccharides, molecules that trigger the immune response, phagocytosis-resistant molecules, and quorum sensing molecules. The main findings from these simulations were that the most important carbon source to produce the virulence factors assessed are tyrosine and glucose and that every virulence factor can be produced by at least one carbon source. These results have been useful for drug and control design, as the specificity of the species for carbon sources was shown.

With all this taken into account, which objective function should be used for metabolic modeling? Which biomass formulation should be used? Or should it be related to pathogenesis or virulence? Or a combination of both? The final answer is in the nature of the biological question or aim to be achieved. A first approach to the model, using the biomass formulation alone, can be used to calibrate the model and assess the normal behavior in standard conditions of *in vitro* culturing. However, if a deeper understanding of the host-pathogen interaction is desired, a pathogenic/virulence focus objective function must be proposed and supported by experimental data. A final comparison between the three results of modeling with: biomass, pathogenic, and a combination of both could give insights into the pathogenic behavior. One example of the improvement of objective function based on experimental data in pathogenesis is in *Ralstonia solanacearum,* where the researchers assessed the trade-off between virulence and proliferation (Peyraud et al., 2016). Another example was proposed by researchers to modify the objective function of *M. tuberculosis* based on proteomics data, successfully improving the predictions under antibiotic stress (Montezano et al., 2015).

In conclusion, metabolic networks may be analyzed by the CBM approach without knowing all the metabolic parameters. The predictions provided by CBM can help uncover the pathogenicity mechanisms in plant pathogenic bacteria. Also, the design of control strategies against pathogens may be done by simulating multiple mutants *in silico* and then testing potential candidates in the laboratory. However, one of the weaknesses of the actual definition of objective function for metabolic studies, is the lack of experimental data to improve and confirm the predictions of the non-model organisms. Thus, unless the utility of top-down approaches for genome-scale modeling is evident, a better effort for obtaining experimental data for non-model organisms is necessary to assess the level of bias of using information of model organisms for non-model ones. Furthermore, other elements must be included in the biomass formulations as metabolic cofactors. These, have an impact in the predictions of growth of different strains on different media, as shown in previous studies (Xavier et al., 2017). Today, the CBM is the standardized approach for conducting metabolic analyses. New methods that complement CBM are being developed and incorporate regulatory, lipidomics, and transcriptomics data. This will certainly help improving the power of the predictions.

PROTEIN-PROTEIN INTERACTION NETWORKS

A fundamental aspect of systems biology is the understanding of the interaction of its components in a holistic way. For networks of proteins, interactions allow the establishment of clusters and routes that proteins develop during a process (Singh et al., 2007). Each of these clusters of interactions describes a function e.g., signal transduction, assembly of the cytoskeleton, protein degradation, etc. (Zhang, 2009).

One of the great advantages that the reconstruction of PPINs provides is the ability to obtain evidence of synergy[2], redundancy[3], re-wiring[4], robustness[5], and even evolutionary processes (Sun et al., 2012). For example, the analysis of disturbance (where individual proteins are eliminated from the network) applied to a PPIN helps identify critical proteins in the system (Yadav and Babu, 2012). In addition, it is possible to integrate PPIN with other kinds of networks (for example regulatory and metabolic networks) or information to improve the understanding of microorganisms (Gligorijević and Pržulj, 2015). Finally, experimental techniques may be used to improve the reconstruction of the PPIN or to validate specific protein-protein interactions. The main experimental techniques used are shown in **Table 3**. A good example of the utility of high-throughput experimental techniques for PPIN reconstruction in plant pathogens is the Yeast Two-Hybrid (Y2H) system. In this study, the interaction between *A. thaliana* and three pathogens: *P. syringae*, *Hyaloperonospora arabidopsidis*, and *Golovinomyces orontii* (Weßling et al., 2014) were assessed. Importantly, the researchers found *Arabidopsis* target elements shared by the three pathogens, highlighting the importance of a few hubs in plants that can be targeted by pathogenicity weapons of the microorganism. This highlights the relevance of the integration of experimental techniques in pathogenicity studies.

In the following section, we will discuss some approaches for the analyses of PPIN in the context of pathogenicity interactions. The concepts used for characterizing and comparing networks were previously defined (**Table 1** and **Box 4**).

Computational Methods in PPIN for Pathogenic Interaction Studies

Among the multiple computational analyses that can be performed for the reconstruction of PPIN and prediction of interactions (**Table 4**), we will focus on phylogenetic methods, used in bacterial pathogens (Albert, 2007). We will also discuss the importance of modeling the dynamics of PPINs and how PPIN can be used for gaining insights into the meaning of pathogenicity. The reader can review other methods for PPINs reconstruction elsewhere (Dyer et al., 2007; Zahiri et al., 2013).

Phylogenetic Methods: Orthologous Domains or Genes

A first methodological approach within PPIN consists of the identification of interacting proteins based on orthologous genes that are known to interact (He et al., 2008). For this approach, databases of interactions from well-characterized organisms such as *Homo sapiens, E. coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster,* can be used. He et al. (2008) used these databases for the prediction of protein-protein interactions for *Magnaporthe grisea*, a pathogenic fungus that produces rice blast disease. In this study, they identified orthologous genes corresponding to proteins that are known to interact using databases from *E. coli*, *S. cerevisae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*. They obtained a network of around 3,000 proteins for *M. grisea*. Among these, 40 seemed to be hubs that showed a high network degree. All the interactions were validated through *in silico* approaches and authors found possible pathogenic clusters involved in infection, such as phosphorus metabolism, chromatin silencing, and ion transport. This study highlights the importance of the network approach for predicting interactions where no previous information for the organism is available.

A second approach uses different protein features related to known protein-protein interactions: a motif, a domain, or a tridimensional structure. Then, these features are used to predict

[2] **Synergy**: union of two or more processes or paths that generate new process or biological properties.

[3] **Redundancy**: repetition of process or elements that serve as a functional reserve in case of failure.

[4] **Rewiring**: change in the association of biological entities that can vary along the time for improvement of system efficiency.

[5] **Robustness**: capacity of the biological system to recover from perturbations conserving the equilibrium of the system.

TABLE 3 | Main experimental techniques used for reconstruction or validation of protein-protein interaction networks.

Technique	Large scale implementation	Binary interaction or complex	Advantage	Disadvantage	Organisms	References
Y2H - Yeast two hybrid	+++	B	No antibody required	Elevated rate of false-positives; Nuclear localization of proteins	*Francisella tularensis; Blumeria Graminis; Pseudomonas syringae; Hyaloperonospora arabidopsidis; Golovinomyces orontii*	Weßling et al., 2014; Wallqvist et al., 2015; Pennington et al., 2016
PCA - Protein-fragment complementation Assay	++	C	Interaction with membrane proteins	Works better with small monomeric proteins	*Vibrio cholerae; Escherichia coli*	Ozawa et al., 2001; Hatzios et al., 2012
FRET - Förster resonance energy transfer	+	B	Reversible interaction	Decreased sensibility; Photobleaching	*Hordeum vulgare*	Bhat et al., 2005
BiFC - Bimolecular fluorescence complementation	+++	B	Used for localization in living cells	Detection of weakly associated proteins	*Agrobacterium tumefaciens*	Lacroix et al., 2005
TAP - Tandem affinity purification-mass spectroscopy	+	C	Accurate and efficient for multiprotein complex	High experimental effort and extensive data analysis	*Measles morbillivirus; Candida albicans*	Kaneko et al., 2004; Komarova et al., 2011
Protein array	++	C	Highly specific recognition	Needs a set of labeled proteins	*Staphylococcus Aureus*	Scietti et al., 2016
Pull - down	+++	C	Medium level of standardization	Protein GST fusion may cause sterical hindrance	*Streptococcus suis*	Li et al., 2016
Phage display	+++	C	Great diversity of variant proteins that can be represented in a phage library	Post-translational modifications; selection condition of library	*Helicobacter pylori*	Jonsson et al., 2004

TABLE 4 | Computational methods for prediction of protein-protein interaction.

Technique	Algorithms	Strengths	Weaknesses	Organism	Reference
Phylogenetic	Cluster analysis, maximum likelihood, maximum parsimony, Bayesian inference	Provides information of selective environmental pressure	Difficult to estimate divergence of proteins	*H. pylori, P. falciparum*	Ratmann et al., 2007
Machine learning	Random forest, decision tree, k-nearest neighbors, bayesian, Neural networks, support vector machine	Simple to understand, accurate	Dependent of parameter settings and features, black-box predictor, large data set for training	*Vibrio cholerae, P. aeruginosa*	Nanni et al., 2012; Ehrenberger et al., 2015
Data mining	Named entity recognition, ID3, Computational of natural language processing, C4.5	Fast and process large volumes of information, good to focused list	It is sensitive to noise, require manually curation	*H. pylori, Campylobacter jejuni*	Bock and Gough, 2003
Topological	*Power-law* degree distribution, clustering coefficient	Common topological characteristics among species (small-world), comparison with random networks	False positives proportional to the size of the network, configuration of protein modules may vary	*E. coli*	Butland et al., 2005; Wuchty, 2006; Sharan et al., 2007
Structure	Shape complementarity, rigid-body docking, heuristic potential	Accurate, good availability of data for primary and secondary structure	Slow development for high throughput methodologies	*E. coli*, *S. typhimurium* and *T. maritima*	Matsuzaki et al., 2014

new interactions (Davis et al., 2007). The predictions of host-pathogen protein interactions have been mostly based on the *S. cerevisiae* interactome[6] [which was reconstructed based on affinity purification/mass spectrometry (Collins et al., 2006)]. This interactome created a reference map which was curated for later studies. For example, Davis et al. (2007) used it to predict interactions of 10 human pathogens, including *Plasmodium* and *Mycobacterium* species, generating a full protocol based on protein domains.

In the case of plant-pathogens, a prediction at the genome-scale was calculated for *A. thaliana* and *P. syringae*. This was done through two methodologies based on domains and interolog[7], generating more than 85,000 interactions, of which 11,000 were shared by the two methodologies (Sahu et al., 2014).

Despite the power of phylogenetic methods, they can be largely affected by the number of genomes used and the quality of their assembly and annotation. Therefore, a robust methodology of verification of false positives is necessary to evaluate the accuracy of these methods.

Modeling Dynamic Networks

Protein networks have been presented so far as a mechanism that allows associations to be viewed in a static way. In contrast, the cell performs processes precisely by receiving and emitting signals in a temporal context (Przytycka et al., 2010). The study of dynamic networks aims to identify changes in topology, function, spatial distribution, and information flow, to understand the organism's response to disturbance in function of time.

For example, probabilistic approaches integrate gene expression profiles from different time points and protein interaction data for the reconstruction of more accurate PPIN than the networks that rely only on one time point (Zhang et al., 2016). This probabilistic approach identifies protein complexes better than static methods and localizes the protein complex in their correct time stage at biological level. The authors exemplify this in the case of a protein complex of the Golgi transport system, showing their interaction in a specific point of the time series (Zhang et al., 2016).

Also, integrative strategies (using proteomics, genomics, and transcriptomics) have been generated to observe changes at the level of protein interaction or gene expression, both permanent or transient for the detection of biomarkers of disease progression as reviewed by (Wang et al., 2014b).

Temporary associations generate rapid response mechanisms, vital in defense processes against pathogens. Therefore, dynamic networks could be used for generating models of disease progression, helping in the design of drugs or control strategies (Przytycka et al., 2010).

PPIN in the Context of Host-Pathogen Interactions

We want to point the main use of PPIN approaches in pathogenicity context. First, multiple protein-protein interactions among pathogenicity factors (e.g., effector proteins) and host proteins (based on genome data and information of related species) can be assessed *in silico.* Then, the target of the protein of the pathogen into the host can be predicted, and obtain a network of PPIN of the pathogen and the host. Second, host-pathogen interactions can be assessed through techniques as Y2H or other techniques mentioned at the beginning of this chapter. Then, these experimental data and *in silico* predictions can be used to construct a PPIN of the host-pathogen interaction. This kind of network has a lot of information useful for the biotechnological control of the pathogen. For example, with the

[6] It is appropriate to clarify that the interactome includes the set of interactions that can occur in an organism, usually but not always, represented by protein-protein interaction.

[7] Interactions that are conserved among pairs of proteins that are present in descent-related species.

help of network metrics (**Table 1**) such as clustering coefficients or network degree, hubs of susceptibility in the host can be detected. Finally, it is highly recommended to experimentally confirm the candidates by more precise techniques such as CoIP.

Signaling Networks: A Special Case of PPIN

Signaling is a series of chemical and/or energetic transmissions from an external stimulus to the cell. Signaling networks reconstruct the interaction path of the signal-carrying elements (usually proteins) to the organelle that requires the decision of maintaining or changing a state of homeostasis (Cho et al., 2015). These networks are also represented in a directed manner (**Figure 1A**) and with highly conserved and specific topologies. Usually, these types of networks also include transcription factors and PPIN to reconstruct the signaling cascade. The most likely edges of signal conduction are also weighted by the strongest or most reliable directed path (Cho et al., 2015). From the computational point of view, the weighting of the vertices constitutes a great challenge due to the inherent subjectivity of this process. Punctuation methodologies have been proposed by defining a probability (Liu and Zhao, 2004).

Kim et al. (2014) discuss the robustness and modularity of an immunity network, specifically that of *A. thaliana* under a pathogen attack, investigating the changes of the plant-immunity process called pattern-triggered immunity. They constructed a dynamic model of a signaling network by evaluating the determination of certain plant hormones against the immune challenge, to evaluate their predictive power. They found that the hormone ethylene increases the robustness of the system by inhibiting the jasmonate pathway. With this, they could conclude that the network is able to grade the level of the response to a given pathogen.

REGULATORY NETWORKS

Regulatory networks represent the relationship between genes and regulatory proteins that lead to the expression or suppression of certain genes. The graphs of regulatory networks are represented in a directed way (**Figure 1A**), trying to capture a series of events that are often consecutive. These networks show highly defined and sometimes hierarchical modules (Lozada-Chávez et al., 2006).

Regulatory networks are highly dependent on the environmental conditions, the cell type that is being studied, and the developmental stages of the organism. Due to the nature of this type of networks, mechanisms of control and modulation generally given by transcription factors need to be considered (Lee, 2002). Moreover, these networks may represent protein-DNA interaction. Thus, they may be easily integrated into protein-interaction networks and metabolic networks.

Because of the large amount of information that is possible to integrate to these networks, multiple approaches have been implemented, based on different sources of information (Marbach et al., 2012). **Table 5** summarize some of the methods used for the reconstruction of regulatory networks. For instance, in a report on the plant pathogen *Xanthomonas axonopodis* pv. *citri,* researchers used microarrays and mutants to decipher the role of two proteins, HrpX and HrpG, in the global control of the virulence process (Guo et al., 2011) and proposed a regulatory model. Also, Seo and collaborators used analysis of gene expression profiles and ChIP-chip experiments to uncover the main transcriptional architecture and regulatory features of *K. pneumoniae* (Seo et al., 2012).

Finally, transcriptional reprogramming is a mechanism of great importance in the control of pathogenicity. Consequently, the reconstruction of regulatory networks derived from temporal series of gene expression data, available in public repositories (Marbach et al., 2012), could be used to predict the response of the pathogen to host defense or antibiotic treatment. Adding promoter regions and functional annotations can help improve this type of network and highlight key components in pathogenicity and evolution of resistance.

NETWORKS, EVOLUTION, AND PATHOGENICITY

Evolution of Network Topology and Distribution of Fluxes

The comparative analysis of networks is a powerful tool that allows understanding the evolutionary relationships among organisms. Furthermore, it allows scientists to decipher the evolution of cell processes such as pathogenicity and adaptation to life on a host. In the context of metabolic networks, three main characteristics can be compared: the similarity of their components, their topology or organization, and the distribution of fluxes. Some studies that are reviewed here show several principles of the evolution of networks in pathogenic bacteria. We would like to highlight two of them: (i) highly connected elements of the network are highly conserved and (ii) in a changing environment, the organism will favor one functional objective at the expense of others.

As stated in the first principle mentioned above, the organization of the networks reflects the evolutionary conservation of its components. Some studies have shown the positive correlation between connectivity of proteins and their degree of conservation (Butland et al., 2005). The organization of the core (shared pathways) and the specific networks are related to the lifestyle of the organism. Regardless of the pathway, the highly-connected enzymes or other elements (regulatory modules and protein interactions) in the network are highly conserved. Furthermore, a scale-free network is vulnerable to the removal of the highly-connected proteins (hubs) but not to the deletion of the less connected proteins. The modularity of the networks reflects the lifestyles of the organisms, as will be discussed in the next section (Butland et al., 2005; Kreimer et al., 2008).

Concerning the second principle, while today we have a better understanding of the way networks are organized or their topology, the evolution and the distribution of fluxes through metabolism have been less studied. Schuetz et al. (2012) compared the evolution of metabolism in microorganisms

TABLE 5 | Methods for reconstruction of regulatory networks.

Approaches	Highlights	Challenges	Organisms studied	References
Differential equations	Network dynamic over time, regulation and optimization of function	High computational demanding, complex parameter optimization	*Mus musculus*, *Candida albicans*	Linde et al., 2015
Boolean	Switch-like behavior, efficient and easy interpretation	Only two states, good in small networks, Only synchronous interactions	*H. pylori*	Franke et al., 2008
Bayesian*	Robust to deal of disturbances, integrated knowledge to increase the support	Non-dynamical, high computational cost, often used a hybrid method to increase the accuracy	*E. coli*	Yang et al., 2011
Neural networks	Allows continuous variables over time, very sensitive for regulated systems, noise-resistant	Computational complex, difficult for training, need a lot of input data	*Caulobacter crescentus, E. coli, Bacillus subtilis*	Yaghoobi et al., 2012; Umarov and Solovyev, 2017
State space model	High computational efficiency, probabilistic framework to simulate the network, determines an optimal threshold value	There are no learning steps	*Saccharomyces cerevisiae, Aspergillus fumigatu*	Do et al., 2009; Koh et al., 2009

**To counteract the stationary problem of Bayesian networks, The dynamic Bayesian network approach was developed.*

to the Pareto optimality. The Pareto optimality or Pareto efficiency is an economic concept stating that one's utility will increase only if someone else's utility diminishes (Sen, 1993). Therefore, in a changing environment, an organism faces a series of trade-offs; the optimality of an objective will be at the expense of another (ATP balance, growth rate, or minimization of fluxes; Schuetz et al., 2012). For example, an organism cannot be optimally adapted to growth in aerobic conditions and anaerobic at the same time. More importantly, authors found a deviance of the metabolism's operation of some mutants from the Pareto surface, which support the author's hypothesis that organisms maintain some space from optimality as evolutionary adaptation under changing environments (Schuetz et al., 2012). Thus, evolution favors flux distributions that minimize adjustments to the new conditions (Schuetz et al., 2012).

Comparative Studies of Networks

Network comparisons between different organisms to study their evolution can be performed with different methods. Some methods compare the contents of the network (e.g., similarity in enzymes, individual pathways, or the whole repertoire) while others compare their structure. We will revise some of these methods mentioning their differences and some of their applications.

The first set of methods calculates indices of similarity or distance between networks, by calculating the similarity or distance between the network components (enzymes, transcription factors, or any other sequence used to construct the network). The similarity between proteins can be simply obtained by their sequence or structure similarity but also by the similarity between the EC (Enzyme Classification) numbers of the corresponding reactions, in the case of metabolic networks (IUBMB. Nomenclature Committee of The International Union of Biochemistry and Molecular Biology, 1992; Heymans and Singh, 2003).

Other methods use the information of the structure of the networks. Forst and Schulten (2001) used sequence similarity combined with information of the corresponding network. They defined the distance between pathways based on all the comprising elements that share the same functional role. In the simplest pathway, the elements of a functional role are the enzyme and its substrate and they can be compared by traditional sequence comparison analysis, if the latter is a protein.

Heymans and Singh (2003) proposed to combine both measures, similarity of the components and network structure using local graph similarity. The graph similarity is calculated on enzyme subsets where all the information contained within the pathways, except for the enzymes, is deleted and the simplified subset is then compared (Heymans and Singh, 2003). However, this method applies to individual pathways and a more inclusive approach was proposed by Forst et al. (2006) in a study where the whole metabolic networks are compared (Heymans and Singh, 2003; Forst et al., 2006).

The fourth set of analyses studies differences in the components of the networks; basically, they compare the insertion or deletion of components in a network. These approaches allow the understanding of the adaptation of organisms to new niches. In a network, two types of pathways can be identified, the essential, present in all organisms, and the non-essential, which are under continuous evolution and are specific to the organism's lifestyle (Mithani et al., 2010). In the Reaction correlation analysis (Mithani et al., 2010, 2011) a Euclidean distance is calculated based on the absence or presence of the reactions in different individuals or strains. In the "all but one analysis" included in the software Rahnuma (Mithani et al., 2009), and then redefined by Mithani et al. (2010), the user can identify pathways and reactions present in some organisms but absent in others. The identification of a core network leads to the construction of an Ancestral Network, a network comprising the reactions present in all species and the definition of species-specific networks (Mithani et al., 2010).

Therefore, a Bayesian model like the one proposed by Mithani et al. (2010) for the study of network evolution allows for the identification of regions of the network under selective pressures, most probably involved in pathogenicity processes. In a study combining different approaches of network evolution analysis, Mithani et al. (2011) showed how these comparative analyses lead to the understanding of the evolution and adaptation strategies of a set of related organisms, some pathogenic and other nonpathogenic. For example, according to the ancestral network reconstruction, it has been suggested that the ancestral pseudomonad was saprotrophic from which more specialized pathogens evolved (Mithani et al., 2010, 2011).

The fifth set of analyses compare the topological features of metabolic networks, especially modularity, for more than 300 bacterial species (Kreimer et al., 2008). These analyses permit studying evolution at a broader phylogenetic scale and relate network characteristics with environmental cues. One of the main results of Kreimer and collaborators was that the environmental factors influence network modularity. Also, symbionts and pathogens show lower modularity while the free-living and less niche-specific bacteria show higher modularity (Kreimer et al., 2008). Moreover, endosymbiotic organisms living in nutrient-restricted niches show both smaller networks and less modularity, losing specific fast-evolving pathways (Kreimer et al., 2008). In this context, modularity is interpreted as subset of functionally related and highly connected reactions or pathways. Thus, pathogens and symbionts have a lower number of modules and connections because they are expected to use the external pathways from the host for its own benefit.

Network Evolution and Pathogenicity

The comparative genomic studies reviewed here take advantage of a higher order of organization based on the structure and properties of the molecular level network-based models. These models allow stating additional hypotheses for the evolution of bacterial pathogens. However, studies based on molecular network models have important limitations as do other comparative genomic studies. Missing data is probably the major drawback, for example on the directionality and kinetic parameters of the reactions.

The study of network evolution will help in the understanding of pathogenicity and in the processes of adaptation of pathogens to new or old hosts. Especially, the organization of orphan genes, the species-specific or pathogen-specific genes, and their connections to the core network will help achieve this goal. New genes arise by different processes: exon shuffling, gene duplication, retrotransposition, mobile elements, lateral gene transfer, *de novo*, and a combination of these mechanisms (Long et al., 2003). Once generated, both duplicated and novel genes are less connected at the beginning, however, the rewiring process differs between these two (Capra et al., 2010). In the case of the pathogenicity-related genes, it is argued that they will always occupy peripheral positions in the networks (Kholodenko et al., 2012), a result expected due to their fast-evolving rates.

The study of the rewiring process of recently evolving genes may be helpful in the pathogenicity studies, given that the rewiring process occurs not only inside the cell but also with its interactors (host or pathogen). In a recent study, it was shown that effector proteins from phylogenetically distant organisms converge to and target highly connected hubs of the immune plant system (Mukhtar et al., 2011; Kholodenko et al., 2012). Thus, this mechanism of host-pathogen interaction could help in the prediction of evolving paths in the pathogen as response to drug or pesticide control (in human and plant pathogens respectively), and therefore partially solve the problem of resistance in pathogens subject of pathogenicity control.

CONCLUSIONS

We have reviewed the metabolic, protein-protein and regulatory networks that have helped understanding disease, mechanisms of pathogenesis and virulence, as well as interactions between bacteria and their hosts.

All types of networks, used for prediction purposes, have both strengths and weaknesses, and provide different types of biological information **Table 6**. Also, we showed how topological and other mathematical approaches can be used to analyze every type of network. For example, CBM, which does not rely on the complete knowledge of the kinetic constants, serves as a useful approach for metabolic analyses in pathogenic bacteria. In contrast, the Boolean analysis of regulatory networks, which relies only on topological features of the network architecture, provides useful information about pathogenic mechanisms. Thus, the different mechanisms of pathogenicity, disease, and virulence can be uncovered by network approaches. However, a strong feedback between the information derived from experimental procedures and computational models should be progressively more relevant and important to improve the conclusions of the models and provide new biological hypotheses.

The systems biology approach can be used to design control strategies of the pathogen. For example, bactericides target important regulators or proteins of the pathogen, identified on *in silico* studies. In the case of regulatory networks, two of the most important aspects related to pathogens are the robustness of the network to random changes and its stability through time. This has been made evident by the high degree of fitness that successful pathogens possess. Pathogens share elements linked to pathogenicity that have simultaneous and/or complementary actions as redundant mechanisms in the event of detection by the host. The robustness is a consequence of the wired redundancy of the gene-regulator interactions, especially in the genes encoding for hub proteins. It can be inferred that the evolutionary forces have shaped and constrained the most important regulatory pathways involved in disease, pathogenicity, and virulence of bacteria. Therefore, genes within pathways that improve the fitness of the pathogen are positively selected, increasing the degree of wiring of these specific mechanisms. These genes are promising targets for bacterial control.

In the case of protein-protein interactions, new methodologies and approaches have emerged from structural, functional and computational knowledge. Studies have focused on the functional role of proteins in disease-related processes, significantly

TABLE 6 | Summary of networks for the study of host-pathogen interaction.

Networks	Experimental data	Mathematical and computational approaches	Objective
Regulatory	Genomics; Transcriptomics; Transcription Start Site (5′-RACE); Binding sites global regulators (ChIP-chip)	Boolean; network analysis	Dynamic of regulation of genes involved in virulence and pathogenicity
Metabolic	Genomics; transcriptomics; Metabolomics; Phenotype microarrays; C13 labeling	Constraint-based modeling; elementary flux mode analysis; pathway enrichment analysis; network analysis	Metabolic capabilities; genes related with virulence and pathogenicity
Protein-protein interaction	Y2H; PCA; BiFC; Protein arrays; Pull down; Phage display	Phylogenetic methods; dynamical networks; machine learning	Identification of hubs involved in virulence and pathogenicity; Determination of interaction between proteins related with signaling and regulatory cascades
Signaling and regulatory	Transcriptomics; Fusion assays (LacZ reporter); Adherence assay; Biofilm formation (fluorescence)	Boolean; network analysis	Impact of sensors in regulation of virulence and pathogenesis; Cell-to-cell signaling; biofilm synthesis;
Signaling, regulatory and metabolic	Genomics; metabolomics; transcriptomics	Constraint-based modeling; boolean model hierarchical layers; network analysis	Model regulatory and metabolic network of QS system

The networks reviewed in this work, the experimental data (mainly at the level of omics), the mathematical and computational approaches applied for every network, and the research objective for the networks studied are summarized.

contributing to the understanding of the role of proteins as mechanistic executors in each of the physiological stages of infection. Thus, signaling pathways or hubs that are susceptible to be blocked to prevent the development of a given disease could be detected and be used to design control strategies of the pathogen. One of these strategies starts from the analysis of domains or contact surfaces allowing to establish interactomes *in silico* and develop mimetic or decoy proteins.

We have shown that regulatory, metabolic, and protein-protein interaction network systems are tightly interconnected, and each of them depends on the others. In the future, we expect that more studies center their efforts into coupled systems using different computational and mathematical approaches with the support of several experimental techniques and approaches (as much targeted to specific genes and mechanisms as supporting high-throughput data analysis). For example, the gene essentiality analysis is important in the context of regulatory networks, where deletion of genes impact molecular networks at the level of protein interactions, signaling cascades, and the metabolic phenotype. Therefore, this analysis constitutes a powerful approach for searching for genetic targets for the design of control strategies against pathogens.

Another example of the inference power of coupled systems is the relationship between the genotype and the phenotype that is reflected in metabolic and protein networks linked to regulatory and signaling networks. It is the convergence of systems, through the switching of the distinct metabolic pathways mediated by regulation of the genes and signaling cascades, that determines the defense and attack mechanisms of the pathogen. The hubs at the level of the regulatory system play an important role in the control of pathogenicity, since a global regulator of pathogenicity can control several genes within a pathogenicity module. Subsequently, the downstream cascade of genes can up or down-regulate several other genes involved in metabolism and other functions. The result is the expression of a metabolic phenotype that serves as a coordinated attack or defense system. Thus, the study of regulatory, signaling and metabolic interactions through a multiscale modeling approach will provide promising results related to pathogenicity and defense mechanisms.

In systems biology, we will see an important improvement of the evolutionary analyses performed on the networks. The incorporation of a genetic population frame is urgently needed to help to understand the pathogenic mechanisms of host-pathogen interactions. A way to accomplish this is through the establishment of relationships between genetic variation of the genes associated with the enzymes and proteins and the properties of the networks to explain this variation in spatial and evolutionary terms in a system context. Ultimately, the host-pathogen relationships are governed by evolutionary forces acting in time and space of the whole biological system.

The evolutionary studies supported by systems biology can help to solve important questions related to pathogenicity as the emergence of specific pathogens and their relationship with non-pathogens. The processes of interaction among species over millions of years have largely been influenced by domestication. This has generated changes among the connections of the elements of the immune system (rewiring). As a result, selection pressures have varied, favoring, in some cases, a non-specific pathogen to infect a given host. This process can be modeled through networks, by reconstructing the routes or proteins of ancestral and/or non-domesticated species and comparing with the present ones to observe the changes in connections among the elements.

From the evolutionary point of view, networks can also demonstrate the molecular changes that have occurred during pathogen interactions. From the hypothesis of arms race processes, new perspectives have been generated that can fill the gaps, such as that proposed by Cook et al. (2015), which provides a view of the host-pathogen interaction, related to mutualism and parasitic symbiosis as initial stages of co-evolution. With the above, we could rethink the approximation strategies and how we understand the interaction of what is considered pathogenic, and how biological networks can drive to new hypothesis through the integration of enormous amount of information.

Finally, the comparisons between pathogens and non-pathogens in an evolutionary context, where there are conserved and divergent features among the different strains and species, can serve to design control strategies and to help to improve the understanding of pathogenicity mechanisms.

In this review, we tried to describe different methodologies to solve biological questions using the networks, giving an overview of the available mathematical approaches. As a growing discipline, network analysis in systems biology still has challenges that must be overcome and must be considered when generating new hypotheses. Some of the challenges that need to be addressed are:

I. At the metabolic level, the objective function should be redefined in a context of host-pathogen relationships (xanthan is a good example; other pathogenic factors can be modeled in the same way).
II. Protein-protein interaction prediction methodologies must have a large amount of data as a basis for prediction.
III. The reconstruction of the regulatory networks still represents experimental limitations since a high amount of data are needed such as time series, gene deletions or biological samples.
IV. The evolutionary forces acting on the networks should be mathematical and computational implemented; not only to compare between different networks of the same species or genus, but also to differentiate among genetic drift, genetic flow and other evolutionary forces.
V. The experimental information on non-model pathogens, especially the high-throughput data must be increased for feeding the computational models and for comparison purposes.

Confronting these challenges will bring the study of pathogenic mechanisms and relationships to a next level. Without doubt, network analysis in systems biology will appear as an essential discipline used in every molecular laboratory that studies host-pathogen interactions and, we will see a burst of user-friendly software in network biology designed for experimental biologist to fulfill this necessity.

AUTHOR CONTRIBUTIONS

DB, CA, AB, GD, SR: Developed and wrote the manuscript; GD, SR, AB: Guided and assisted in writing the manuscript; All authors read and approved the final manuscript.

FUNDING

Faculty of Sciences of the Universidad de Los Andes and the Administrative Department of Science and Technology in Colombia.

ACKNOWLEDGMENTS

We want to thank Faculty of Sciences of the Universidad de Los Andes for all the funding that made this review possible. We also want to thank the Administrative Department of Science and Technology in Colombia (Colciencias) for the award under contract number 0794-2013 for support to CA.

REFERENCES

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., Nielsen, J., et al. (2013). The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput. Biol.* 9:e1002980. doi: 10.1371/journal.pcbi.1002980

Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *Plant Cell* 19, 3327–3338. doi: 10.1105/tpc.107.054700

Aloy, P., and Russell, R. B. (2004). Taking the mystery out of biological networks. *EMBO Rep.* 5, 349–350. doi: 10.1038/sj.embor.7400129

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.

Bartell, J. A., Yen, P., Varga, J. J., Goldberg, J. B., and Papin, J. A. (2014). Comparative metabolic systems analysis of pathogenic *Burkholderia*. *J. Bacteriol.* 196, 210–226. doi: 10.1128/JB.00997-13

Bhat, R. A., Miklis, M., Schmelzer, E., Schulze-Lefert, P., and Panstruga, R. (2005). Recruitment and interaction dynamics of plant penetration resistance components in a plasma membrane microdomain. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3135–3140. doi: 10.1073/pnas.0500012102

Bock, J. R., and Gough, D. A. (2003). Whole-proteome interaction mining. *Bioinformatics* 19, 125–135. doi: 10.1093/bioinformatics/19.1.125

Bonde, B. K., Beste, D. J. V., Laing, E., Kierzek, A. M., and McFadden, J. (2011). Differential Producibility Analysis (DPA) of transcriptomic data with metabolic networks: deconstructing the metabolic response of *M. tuberculosis*. *PLoS Comput. Biol.* 7:e1002060. doi: 10.1371/journal.pcbi.1002060

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., and Jamshidi, N. (2010). Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Mol. Syst. Biol.* 6:422. doi: 10.1038/msb.2010.68

Bruggeman, F. J., and Westerhoff, H. V. (2007). The nature of systems biology. *Trends Microbiol.* 15, 45–50. doi: 10.1016/j.tim.2006.11.003

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., et al. (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, 433, 531–537. doi: 10.1038/nature03239

Capra, J. A., Pollard, K. S., and Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol.* 11:R127. doi: 10.1186/gb-2010-11-12-r127

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 42, D459–D471. doi: 10.1093/nar/gkt1103

Chandrasekaran, S., and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17845–17850. doi: 10.1073/pnas.1005139107

Charusanti, P., Chauhan, S., McAteer, K., Lerman, J. A., Hyduke, D. R., Motin, V. L., et al. (2011). An experimentally-supported genome-scale metabolic network reconstruction for *Yersinia pestis* CO92. *BMC Syst. Biol.* 5:163. doi: 10.1186/1752-0509-5-163

Cho, Y. R., Xin, Y., and Speegle, G. (2015). P-finder: reconstruction of signaling networks from protein-protein interactions and GO annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 309–321. doi: 10.1109/TCBB.2014.2355216

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* 31, 6633–6639. doi: 10.1093/nar/gkg847

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., et al. (2009). Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* 5:e1000489. doi: 10.1371/journal.pcbi.1000489

Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C. P., et al. (2006). Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 6, 439–450. doi: 10.1074/mcp.M600381-MCP200

Cook, D. E., Mesarich, C. H., and Thomma, B. P. H. J. (2015). Understanding plant immunity as a surveillance system to detect invasion. *Annu. Rev. Phytopathol.* 53, 541–563. doi: 10.1146/annurev-phyto-080614-120114

Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H., and Sali, A. (2007). Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 16, 2585–2596. doi: 10.1110/ps.073228407

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717–729. doi: 10.1038/nrmicro2419

Do, J. H., Yamaguchi, R., and Miyano, S. (2009). Exploring temporal transcription regulation structure of *Aspergillus fumigatus* in heat shock by state space model. *BMC Genomics* 10:306. doi: 10.1186/1471-2164-10-306

Duan, G., Christian, N., Schwachtje, J., Walther, D., and Ebenhöh, O. (2013). The metabolic interplay between plants and phytopathogens. *Metabolites* 3, 1–23. doi: 10.3390/metabo3010001

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 23, i159–i166. doi: 10.1093/bioinformatics/btm208

Ehrenberger, T., Cantley, L. C., and Yaffe, M. B. (2015). Computational prediction of protein-protein interactions. *Methods Mol. Biol.* 1278, 57–75. doi: 10.1007/978-1-4939-2425-7_4

Fong, N. L., Lerman, J. A., Lam, I., Palsson, B. O., and Charusanti, P. (2013). Reconciling a *Salmonella enterica* metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal ppc deletion mutant. *FEMS Microbiol. Lett.* 342, 62–69. doi: 10.1111/1574-6968.12109

Forst, C. V., Flamm, C., Hofacker, I. L., and Stadler, P. F. (2006). Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics* 7:67. doi: 10.1186/1471-2105-7-67

Forst, C. V., and Schulten, K. (2001). Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* 52, 471–489. doi: 10.1007/s002390010178

Francke, C., Siezen, R. J., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* 13, 550–558. doi: 10.1016/j.tim.2005.09.001

Franke, R., Müller, M., Wundrack, N., Gilles, E.-D., Klamt, S., Kähne, T., et al. (2008). Host-pathogen systems biology: logical modelling of hepatocyte growth factor and *Helicobacter pylori* induced c-Met signal transduction. *BMC Syst. Biol.* 2:4. doi: 10.1186/1752-0509-2-4

Ganter, M., Bernard, T., Moretti, S., Stelling, and Pagni, M. (2013). MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, 29, 815–816. doi: 10.1093/bioinformatics/btt036

Gligorijević, V., and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface* 12:20150571. doi: 10.1098/rsif.2015.0571

Goh, K. I., Kahng, B., and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.* 87(27 Pt 1):278701. doi: 10.1103/PhysRevLett.87.278701

Guo, Y., Figueiredo, F., Jones, J., and Wang, N. (2011). HrpG and HrpX play global roles in coordinating different virulence traits of *Xanthomonas axonopodis* pv. *citri*. *Mol. Plant Microbe Interact.* 24, 649–661. doi: 10.1094/MPMI-09-10-0209

Hatzios, S. K., Ringgaard, S., Davis, B. M., and Waldor, M. K. (2012). Studies of dynamic protein-protein interactions in bacteria using *Renilla* luciferase complementation are undermined by nonspecific enzyme inhibition. *PLoS ONE* 7:e43175. doi: 10.1371/journal.pone.0043175

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* 2:e100. doi: 10.1371/journal.pcbi.0020100

He, F., Zhang, Y., Chen, H., Zhang, Z., and Peng, Y.-L. (2008). The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genomics* 9:519. doi: 10.1186/1471-2164-9-519

Heymans, M., and Singh, A. K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* 19, i138–i146. doi: 10.1093/bioinformatics/btg1018

Horn, F., and Jackson, R. (1972). General mass action kinetics. *Arch. Ration. Mech. Anal.* 47, 81–116. doi: 10.1007/BF00251225

Hou, Z., Jiang, P., Swanson, S. A., Elwell, A. L., Nguyen, B. K. S., Bolin, J. M., et al. (2015). A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* 5:9570. doi: 10.1038/srep09570

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.* 13, 281–291. doi: 10.1093/bib/bbr049

IUBMB. Nomenclature Committee of The International Union of Biochemistry, and Molecular Biology (1992). *Enzyme Nomenclature 1992 : Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. San Diego, CA: Academic Press.

Jönsson, K., Guo, B. P., and Mekalanos, J. J. (2004). Molecular cloning and characterization of two *Helicobacter pylori* genes coding for plasminogen-binding proteins. *Proc. Natl. Acad. Sci.* 101, 1852–1857. doi: 10.1073/pnas.0307329101

Kalkatawi, M., Alam, I., and Bajic, V. B. (2015). BEACON: automated tool for Bacterial GEnome Annotation ComparisON. *BMC Genomics*, 16:616. doi: 10.1186/s12864-015-1826-4

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kaneko, A., Umeyama, T., Hanaoka, N., Monk, B. C., Uehara, Y., and Niimi, M. (2004). Tandem affinity purification of the *Candida albicans* septin protein complex. *Yeast* 21, 1025–1033. doi: 10.1002/yea.1147

Karr, J. R. R., Sanghvi, J. C. C., Macklin, D. N. N., Gutschow, M. V. V., Jacobs, J. M. M., Bolival, B., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. doi: 10.1016/j.cell.2012.05.044

Képès, F. (2007). *Biological Networks, Vol. 3.* Singapore: World Scientific.

Kholodenko, B., Yaffe, M. B., and Kolch, W. (2012). Computational approaches for analyzing information flow in biological networks. *Sci. Signal.* 5:re1. doi: 10.1126/scisignal.2002961

Kim, P.-J., Lee, D.-Y., Kim, T. Y., Lee, K. H., Jeong, H., Lee, S. Y., et al. (2007). Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13638–13642. doi: 10.1073/pnas.0703262104

Kim, Y., Tsuda, K., Igarashi, D., Hillmer, R. A., Sakakibara, H., Myers, C. L., et al. (2014). Mechanisms underlying robustness and tunability in a plant immune signaling network. *Cell Host Microbe* 15, 84–94. doi: 10.1016/j.chom.2013.12.002

Koh, C., Wu, F. X., Selvaraj, G., and Kusalik, A. J. (2009). Using a state-space model and location analysis to infer time-delayed regulatory networks. *EURASIP J. Bioinform. Syst. Biol.* 2009:484601. doi: 10.1155/2009/484601

Komarova, A. V., Combredet, C., Meyniel-Schicklin, L., Chapelle, M., Caignard, G., Camadro, J.-M. et al. (2011). Proteomic analysis of virus-host interactions in an infectious context using recombinant viruses. *Mol. Cell. Proteomics* 10:M110.007443. doi: 10.1074/mcp.M110.007443

Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. (2008). The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci.* 105, 6976–6981. doi: 10.1073/pnas.0712149105

Kurubanjerdjit, N., Tsai, J. J. P., and Ng, K. (2012). "Prediction of microRNA-regulated *A. thaliana*-Xcc protein interaction pathways," in *International Conference on Agricultural, Environment and Biological Sciences* (Phuket), 6–9.

Lacroix, B., Vaidya, M., Tzfira, T., and Citovsky, V. (2005). The VirE3 protein of *Agrobacterium* mimics a host cell function required for plant genetic transformation. *EMBO J.* 24, 428–437. doi: 10.1038/sj.emboj.7600524

Lee, T. I. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804. doi: 10.1126/science.1075090

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. *Nat. Rev. Microbiol.* 10, 291–305. doi: 10.1038/nrmicro2737

Li, X., Liu, P., Gan, S., Zhang, C., Zheng, Y., Jiang, Y., et al. (2016). S. *suis* protein Fhb: mechanisms of host-pathogen protein complex formation and bacterial immune evasion of *Streptococcus suis* protein Fhb. *J. Biol. Chem.* 291, 17122–17132. doi: 10.1074/jbc.M116.719443

Liao, Y.-C., Huang, T.-W., Chen, F.-C., Charusanti, P., Hong, J. S. J., Chang, H.-Y., et al. (2011). An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J. Bacteriol.* 193, 1710–1717. doi: 10.1128/JB.01218-10

Linde, J., Schulze, S., Henkel, S. G., and Guthke, R. (2015). Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J.* 14, 346–378. doi: 10.17179/excli2015-168

Liu, Y., and Zhao, H. (2004). A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. *BMC Bioinformatics* 5:158. doi: 10.1186/1471-2105-5-158

Lobel, L., Sigal, N., Borovok, I., Ruppin, E., and Herskovits, A., a. (2012). Integrative genomic analysis identifies isoleucine and CodY as regulators of *Listeria monocytogenes* virulence. *PLoS Genet.* 8:e1002887. doi: 10.1371/journal.pgen.1002887

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875. doi: 10.1038/nrg1204

Lozada-Chávez, I., Janga, S. C., and Collado-Vides, J. (2006). Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 34, 3434–3445. doi: 10.1093/nar/gkl423

Marbach, D., Costello, J. C., Küffner, R., Vega, N. N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* 296, 910–913. doi: 10.1126/science.1065103

Matsuzaki, Y., Ohue, M., Uchikoga, N., and Akiyama, Y. (2014). Protein-protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis. *Protein Pept. Lett.* 21, 790–798. doi: 10.2174/09298665113209990066

McCloskey, D., Palsson, B. Ø., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* 9:661. doi: 10.1038/msb.2013.18

Mithani, A., Hein, J., and Preston, G. M. (2011). Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in Pseudomonas. *Mol. Biol. Evol.* 28, 483–499. doi: 10.1093/molbev/msq213

Mithani, A., Preston, G. M., and Hein, J. (2009). Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics* 25, 1831–1832. doi: 10.1093/bioinformatics/btp269

Mithani, A., Preston, G. M., and Hein, J. (2010). A Bayesian approach to the evolution of metabolic networks on a phylogeny. *PLoS Comput. Biol.* 6:e1000868. doi: 10.1371/journal.pcbi.1000868

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., et al. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20338–20343. doi: 10.1073/pnas.1307797110

Montezano, D., Meek, L., Gupta, R., Bermudez, L. E., and Bermudez, J. C. M. (2015). Flux balance analysis with objective function defined by proteomics data-metabolism of *Mycobacterium tuberculosis* exposed to mefloquine. *PLoS ONE* 10:e0134014 . doi: 10.1371/journal.pone.0134014

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35(Suppl. 2), W182–W185. doi: 10.1093/nar/gkm321

Mukhtar, M. S., Carvunis, A.-R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., et al. (2011). Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* 333, 596–601. doi: 10.1126/science.1203659

Nanni, L., Lumini, A., Gupta, D., and Garg, A. (2012). Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's Pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 467–475. doi: 10.1109/TCBB.2011.117

Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press.

Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5:320. doi: 10.1038/msb.2009.77

Oberhardt, M. A., Puchałka, J., dos Santos, V. A. P. M., and Papin, J. A. (2011). Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput. Biol.* 7:e1001116. doi: 10.1371/journal.pcbi.1001116

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* 9:693. doi: 10.1038/msb.2013.52

Orth, J. D., and Palsson, B. Ø. (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* 107, 403–412. doi: 10.1002/bit.22844

Orth, J., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Ozawa, T., Takeuchi, M., Kaihara, A., Sato, M., and Umezawa, Y. (2001). Protein splicing-based reconstitution of split green fluorescent protein for monitoring protein-protein interactions in bacteria: improved sensitivity and reduced screening time. *Anal. Chem.* 73, 5866–5874. doi: 10.1021/ac010717k

Pennington, H. G., Gheorghe, D. M., Damerum, A., Pliego, C., Spanu, P. D., Cramer, R., et al. (2016). Interactions between the powdery mildew effector BEC1054 and barley proteins identify candidate host targets. *J. Proteome Res.* 15, 826–839. doi: 10.1021/acs.jproteome.5b00732

Perumal, D., Lim, C. S., and Sakharkar, M. K. (2009). A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identification. *Summit Translat. Bioinforma.* 2009, 100–104.

Peyraud, R., Cottret, L., Marmiesse, L., Gouzy, J., and Genin, S. (2016). A resource allocation trade-off between virulence and proliferation drives metabolic versatility in the plant pathogen *Ralstonia solanacearum*. *PLoS Pathog.* 12:e1005939. doi: 10.1371/journal.ppat.1005939

Pinzón, A., Rodriguez,-R. L. M., González, A., Bernal, A., and Restrepo, S. (2011). Targeted metabolic reconstruction: a novel approach for the characterization of plant-pathogen interactions. *Brief. Bioinform.* 12, 151–162. doi: 10.1093/bib/bbq009

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief. Bioinform.* 11, 15–29. doi: 10.1093/bib/bbp057

Puchałka, J., and Kierzek, A. M. (2004). Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophys. J.* 86, 1357–1372. doi: 10.1016/S0006-3495(04)74207-1

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Comput. Biol.* 3:e230. doi: 10.1371/journal.pcbi.0030230

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Reed, J. L., Famili, I., Thiele, I., and Palsson, B. Ø. (2006). Towards multidimensional genome annotation. *Nat. Rev. Genet.* 7, 130–141. doi: 10.1038/nrg1769

Richardson, E. J., and Watson, M. (2013). The automatic annotation of bacterial genomes. *Brief. Bioinform.* 14, 1–12. doi: 10.1093/bib/bbs007

Sahu, S. S., Weirick, T., and Kaundal, R. (2014). Predicting genome-scale Arabidopsis-*Pseudomonas syringae* interactome using domain and interolog-based approaches. *BMC Bioinformatics* 15:S13. doi: 10.1186/1471-2105-15-S11-S13

Savageau, M. A. (1969). Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* 25, 365–369.

Schaadt, N. S., Steinbach, A., Hartmann, R. W., and Helms, V. (2013). Rule-based regulatory and metabolic model for Quorum sensing in *P. aeruginosa*. *BMC Syst. Biol.* 7:81. doi: 10.1186/1752-0509-7-81

Schatschneider, S., Persicke, M., Watt, S. A., Hublik, G., Pühler, A., Niehaus, K., et al. (2013). Establishment, *in silico* analysis, and experimental verification of a large-scale metabolic network of the xanthan producing *Xanthomonas campestris* pv. *campestris* strain B100. *J. Biotechnol.* 167, 123–134. doi: 10.1016/j.jbiotec.2013.01.023

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinform.* 11:213. doi: 10.1186/1471-2105-11-213

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science* 336, 601–604. doi: 10.1126/science.1216882

Schulze, S., Henkel, S. G., Driesch, D., Guthke, R., Linde, J. J., Sebastian, H., et al. (2015). Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Front. Microbiol.* 6:65. doi: 10.3389/fmicb.2015.00065

Scietti, L., Sampieri, K., Pinzuti, I., Bartolini, E., Benucci, B., Liguori, A., et al. (2016). Exploring host-pathogen interactions through genome wide protein microarray analysis. *Sci. Rep.* 6:27996. doi: 10.1038/srep27996

Segrè, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117. doi: 10.1073/pnas.232349399

Sen, A. (1993). Markets and freedoms: achievements and limitations of the market mechanism in promoting individual freedoms. *Oxf. Econ. Pap.* 45, 519–541. doi: 10.1093/oxfordjournals.oep.a042106

Seo, J.-H., Hong, J. S.-J., Kim, D., Cho, B.-K., Huang, T.-W., Tsai, S.-F., et al. (2012). Multiple-omic data analysis of *Klebsiella pneumoniae* MGH 78578 reveals its transcriptional architecture and regulatory features. *BMC Genomics* 13:679. doi: 10.1186/1471-2164-13-679

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3:88. doi: 10.1038/msb4100129

Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on-off minimization of metabolic flux. *Proc. Natl. Acad. Sci. U.S.A.* 102,7695–7700. doi: 10.1073/pnas.0406346102

Singh, R., Xu, J., and Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology. *Res. Comput. Mol. Biol.* 4453, 16–31. doi: 10.1007/978-3-540-71681-5_2

Steinway, S. N., Biggs, M. B., Loughran, T. P., Papin, J. A., and Albert, R. (2015). Inference of network dynamics and metabolic interactions in the gut microbiome. *PLoS Comput. Biol.* 11:e1004338. doi: 10.1371/journal.pcbi.1004338

Sun, M. G. F., Sikora, M., Costanzo, M., Boone, C., and Kim, P. M. (2012). Network evolution: rewiring and signatures of conservation in signaling. *PLoS Comput. Biol.* 8:e1002411. doi: 10.1371/journal.pcbi.1002411

Swainston, N., Smallbone, K., Mendes, P., Kell, D., and Paton, N. (2011). The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J. Integr. Bioinform.* 8:186. doi: 10.1515/jib-2011-186

Thiele, I., and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Biotechnol.* 5, 93–121. doi: 10.1038/nprot.2009.203

Thiele, I., Hyduke, D. R., Steeb, B., Fankam, G., Allen, D. K., Bazzani, S., et al. (2011). A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella Typhimurium* LT2. *BMC Syst. Biol.* 5:8. doi: 10.1186/1752-0509-5-8

Umarov, R. K., and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* 12:e0171410. doi: 10.1371/journal.pone.0171410

Varma, A., and Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 60, 3724–3731.

Vorhölter, F. J., Schneiker, S., Goesmann, A., Krause, L., Bekel, T., Kaiser, O., et al. (2008). The genome of Xanthomonas campestris pv. campestris B100 and its use for the reconstruction of metabolic pathways involved in xanthan biosynthesis. *J. Biotechnol.* 134, 33–45. doi: 10.1016/j.jbiotec.2007.12.013r

Wallqvist, A., Memišević, V., Zavaljevski, N., Pieper, R., Rajagopala, S. V., Kwon, K., et al. (2015). Using host-pathogen protein interactions to identify and characterize *Francisella tularensis* virulence factors. *BMC Genomics* 16:1106. doi: 10.1186/s12864-015-2351-1

Wang, C., Deng, Z.-L. L., Xie, Z.-M. M., Chu, X.-Y. Y., Chang, J.-W. W., Kong, D.-X. X., et al. (2014a). Construction of a genome-scale metabolic network of the plant pathogen *Pectobacterium carotovorum* provides new strategies for bactericide discovery. *FEBS Lett.* 589, 285–294. doi: 10.1016/j.febslet.2014.12.010

Wang, J., Peng, X., Peng, W., and Wu, F.-X. (2014b). Dynamic protein interaction network construction and applications. *Proteomics* 14, 338–352. doi: 10.1002/pmic.201300257

Ward, J. L., Forcat, S., Beckmann, M., Bennett, M., Miller, S. J., Baker, J. M., et al. (2010). The metabolic transition during disease following infection of *Arabidopsis thaliana* by *Pseudomonas syringae* pv. *tomato. Plant J.* 63, 443–457. doi: 10.1111/j.1365-313X.2010.04254.x

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440–442.

Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., et al. (2014). Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* 16, 364–375. doi: 10.1016/j.chom.2014.08.004

Winterbach, W., Mieghem, P., Van, R.einders, M., Wang, H., and de Ridder, D. (2013). Topology of molecular interaction networks. *BMC Syst. Biol.* 7:90. doi: 10.1186/1752-0509-7-90

Wuchty, S. (2006). Topology and weights in a protein domain interaction network - a novel way to predict protein interactions. *BMC Genomics* 7:122. doi: 10.1186/1471-2164-7-122

Xavier, J. C., Patil, K. R., and Rocha, I. (2017). Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab. Eng.* 39, 200–208. doi: 10.1016/j.ymben.2016.12.002

Yadav, G., and Babu, S. (2012). Nexcade: perturbation analysis for complex networks. *PLoS ONE* 7:e41827. doi: 10.1371/journal.pone.0041827

Yaghoobi, H., Haghipour, S., Hamzeiy, H., and Asadi-Khiavi, M. (2012). A review of modeling techniques for genetic regulatory networks. *J. Med. Signals Sens.* 2, 61–70.

Yang, B., Zhang, J., Shang, J., and Li, A. (2011). "A Bayesian network based algorithm for gene regulatory network reconstruction," in *2011 IEEE International Conference on Signal Processing, Communications and Computing* (ICSPCC), 1–22.

Zahiri, J., Bozorgmehr, J. H., and Masoudi-Nejad, A. (2013). Computational prediction of protein–protein interaction networks: algorithms and resources. *Curr. Genomics* 14, 397–414. doi: 10.2174/1389202911314060004

Zanghellini, J., Ruckerbauer, D. E., Hanscho, M., and Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell: properties, calculation and applications. *Biotechnol. J.* 8, 1009–1016. doi: 10.1002/biot.201200269

Zhang, A. (2009). *Protein Interaction Networks - Computational Analysis*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511626593

Zhang, Y., Lin, H., Yang, Z., and Wang, J. (2016). Construction of dynamic probabilistic protein interaction networks for protein complex identification. *BMC Bioinformatics* 17:186. doi: 10.1186/s12859-016-1054-1

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.

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