COMPUTATIONAL METHODS FOR MICROBIOME ANALYSIS

EDITED BY: Joao Carlos Setubal, Jens Stoye and Bas E. Dutilh
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-437-5
DOI 10.3389/978-2-88966-437-5

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

COMPUTATIONAL METHODS FOR MICROBIOME ANALYSIS

Topic Editors:
Joao Carlos Setubal, University of São Paulo, Brazil
Jens Stoye, Bielefeld University, Germany
Bas E. Dutilh, Utrecht University, Netherlands

Citation: Setubal, J. C., Stoye, J., Dutilh, B. E., eds. (2021). Computational Methods for Microbiome Analysis. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-437-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 11 December 2020
doi: 10.3389/fgene.2020.623897

Editorial: Computational Methods for Microbiome Analysis

João C. Setubal[1]*, *Jens Stoye*[2] *and Bas E. Dutilh*[3]

[1] *Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil,* [2] *Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany,* [3] *Theoretical Biology and Bioinformatics, Department of Biology, Science for Life, Utrecht University, Utrecht, Netherlands*

Keywords: bioinformatics, metagenomics, microbiome, microbial ecology, taxonomic classification

Editorial on the Research Topic

Computational Methods for Microbiome Analysis

Microbes play critical roles in the lives of hosts (plants, animals, humans) and in almost any environment one can think of. Gathering microbiome sequence data has become easier and cheaper than ever before, leading to an exponential growth in the amount of such data available for analysis. With this explosion has come a pressing need for sophisticated computational tools that can help make sense of these datasets. Current challenges, such as the complexity of microbiome-host-environment interactions and the large sizes of datasets, make for a fascinating research field.

The goal of this Research Topic was to gather a collection of high-quality original papers on the general theme of computational methods for microbiome analysis. We now present the results, which consist of 13 papers.

Four papers consider amplicon analysis, a popular method for taxonomic classification of mixed microbial samples based e.g., on 16S rRNA gene regions. The paper by Engelmann et al. describes Cascabel, a software pipeline for automated processing and analyzing of massive amounts of amplicon data. Cascabel wraps around existing and well-established tools in the field of amplicon analysis, connecting them by means of a Snakemake workflow, thus allowing for an easy and flexible execution of a common amplicon analysis pipeline. After a workflow is finished, reports are generated also serving as a data provenance description.

Similar in flavor is NG-Tax 2.0, described in the paper by Poncheewin et al. which follows the new amplicon sequencing variant (ASV) approach, where sequencing reads are grouped into ASV clusters of very high similarity in order to sustain as much as possible the true biological variance in the sample at hand. NG-Tax 2.0 performs several steps in a pipeline manner, which includes demultiplexing, read cleaning, ASV clustering, and taxonomic classification.

In a somewhat theoretical study based on computer simulations, Pinna et al. try to answer the question of whether non-contiguous V-regions with paired-end sequencing improve 16S rRNA based taxonomic resolution of microbiomes. They explore the possibility of combining two regions of the 16S rRNA gene for a better classification of the tags and to possibly iron out the weakness of taxonomic resolution of one region by a higher resolved other region. And indeed, the combination of two distant variable regions shows on average 10-20 percent higher accuracy in taxonomic classification—a theoretical potential, however, that still needs to be explored in practice.

Still on the topic of taxonomic classification, Shah et al. present ATLAS, a novel strategy for taxonomic annotation of 16S rRNA sequence data. It has been recognized that 16S amplicon data does not in general allow reliable classification below the genus level. However, 16S sequence data and the accumulated knowledge on the diversity of this marker gene (as present in various 16S databases) may allow reliable classification at the "sub-genus level," meaning classification that would suggest possible species, to the exclusion of others, of the same genus. That is the main

OPEN ACCESS

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

****Correspondence:***
João C. Setubal
setubal@iq.usp.br

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

Received: *30 October 2020*
Accepted: *24 November 2020*
Published: *11 December 2020*

Citation:
Setubal JC, Stoye J and Dutilh BE
(2020) Editorial: Computational
Methods for Microbiome Analysis.
Front. Genet. 11:623897.
doi: 10.3389/fgene.2020.623897

achievement of the ATLAS pipeline, which is therefore a contribution for better use of the large amounts of 16S data already available and yet to come.

Moving on to other subtopics, Dong and Strous present MetaErg, a user-friendly platform to explore the information in complex metagenomic datasets. It facilitates the annotation, visualization, and interpretation of assembled and/or binned shotgun metagenomes by using taxonomic and functional annotation of genes identified in metagenomic contigs or metagenome-assembled genomes, or MAGs. Homology searches may be performed with DIAMOND-blast as well as a collection of HMM-profiles. Moreover, MetaErg allows the incorporation of additional -omics data such as metaproteomics to identify gene expression. The HTML-based output pages allow users to navigate through annotation tables, trees, and sunburst plots to explore their data.

Accurate metagenome assembly and genome binning from short-read data may be confounded by e.g., repeated regions and mobile genetic elements. Chromosomal contact data from meta3C or Hi-C experiments is a promising way to address these challenges. Baudry et al. present MetaTOR, a binning pipeline that uses contact frequencies to reconstruct MAGs from meta3C metagenomic libraries. Application to murine gut metagenomes enabled the recovery of MAGs corresponding to nearly a third of the total assembly data of 20 meta3C libraries, underlining the promise of chromosomal contact data for metagenome-binning and the potential to describe microbial communities with MAGs.

Hester et al. present a new metric for evaluating functional redundancy in metagenomes that they call metabolic overlap (MO). The metric needs annotated MAGs of each environment considered. They observed highest values of MO for aquatic and low pH/high temperature environments, and lowest values in communities associated with animal hosts, in one built/engineered environment, and in soil. It is an excellent example of an analysis method that seeks to unlock the rich information contained in MAGs to help understand competition and cooperation between species

Within the rapidly expanding field of microbiome science, it is becoming difficult to stay up-to-date with the literature on microbe-human interactions, as a lot of new information is being published. Srivastava et al. present EviMass, a new tool to gain information about microbial associations to the human superorganism from literature. Evimass consists of an interactive query system on top of a large database derived from mined microbe-microbe and disease-microbe associations from PubMed abstracts. Thus, by uploading their own microbial interaction data, users can link these associations to information from biomedical literature. Various output formats and statistics are available, allowing researchers to place their microbiome experiments among the wealth of information in the literature.

Gene-targeted assembly is a useful approach to identify and track specific genes in metagenomic datasets. Guo et al. present a benchmark comparing the computational efficiency, sensitivity, specificity, and chimera rate of six existing gene-targeted assembly tools. The authors focused on extracting the universal ribosomal protein rplB and two nitrogen cycle genes, dinitrogenase reductase gene nifH, and nitrite reductase gene nirK from testing datasets consisting of known genomes, synthetic and mock communities, and a large soil shotgun metagenome. They assessed assembly quality as well as computational performance of the tools. Two tools that employ probabilistic graph structures showed the best overall performance.

Metagenomics is providing unprecedented insights into our microbial world. Combined datasets generated by research laboratories around the world are opening up new opportunities to study the macroecological patterns on local-to-global scales. Mascarenhas et al. contributed a valuable and extensive review on the computational methods to investigate the macroecology of microbiomes. They address fundamental aspects of biodiversity, describe macroecological studies in the microbiology field, and stress how spatial and temporal sampling scales should fit the research question of each study. Next they describe methods including taxonomic profiling and co-occurrence networks, identifying keystones, and description of functional patterns. An important part of their review is a discussion of different approaches for predictive modeling which promise new insights in a range of fields.

To investigate the transcriptional activity of the microbial community, metatranscriptomics requires a specific subset of analysis tools. Shakya et al. review computational tools and recent advances in metatranscriptome analysis. Discussing metatranscriptomics studies investigating diverse ecosystems, they highlight the ability of metatranscriptomics to reveal the transcriptional activity of microbial communities with sometimes high resolution. The authors next discuss different bioinformatics tools and workflows including preprocessing, assembly, taxonomic and functional annotation, and differential expression analysis. They envision that the described tools will aid in the analysis of e.g., time series data to reveal the response of microbial communities to perturbations, although benchmarking is still needed.

Much emphasis has been given in the past to "snapshot" analysis of microbial communities, i.e., analysis of a community at a given point in time. However, for many environments, time is a crucial variable for the understanding of its microbial ecology. Hence, time-series sampling becomes an important strategy. This approach requires specialized techniques, which are nicely presented in the primer by Coenen et al. The authors describe several modules (interactive tutorials in R and Matlab) that address several topics in time-series analysis.

Van den Bogert et al. discuss various challenges for bioinformatics and data science in industrial microbiome applications. They review current applications and products in food, cosmetics and health industries. Some of the challenges facing these applications are also mentioned. Recent technological developments in the microbiome field are discussed and suggestions are given for how these developments could be leveraged to address certain challenges.

In sum, we believe the papers presented form a valuable collection for students and researchers working on the exciting and rapidly-growing field of microbiome analysis.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

We would like to thank authors, reviewers, and Frontiers staff for helping make possible this Research Topic collection.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 July 2019
doi: 10.3389/fgene.2019.00653

Can Targeting Non-Contiguous V-Regions With Paired-End Sequencing Improve 16S rRNA-Based Taxonomic Resolution of Microbiomes?: An *In Silico* Evaluation

Nishal Kumar Pinna[†], *Anirban Dutta**[†], *Mohammed Monzoorul Haque and Sharmila S. Mande**

Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune, Maharashtra, India

Background: Next-generation sequencing (NGS) technologies have enabled probing of microbial diversity in different environmental niches with unprecedented sequencing depth. However, due to read-length limitations of popular NGS technologies, 16S amplicon sequencing-based microbiome studies rely on targeting short stretches of the 16S rRNA gene encompassing a selection of variable (V) regions. In most cases, such a short stretch constitutes a single V-region or a couple of V-regions placed adjacent to each other on the 16S rRNA gene. Given that different V-regions have different resolving ability with respect to various taxonomic groups, selecting the optimal V-region (or a combination thereof) remains a challenge.

Methods: The accuracy of taxonomic profiles generated from sequences encompassing 1) individual V-regions, 2) adjacent V-regions, and 3) pairs of non-contiguous V-regions were assessed and compared. Subsequently, the discriminating capability of different V-regions with respect to different taxonomic lineages was assessed. The possibility of using paired-end sequencing protocols to target combinations of non-adjacent V-regions was finally evaluated with respect to the utility of such an experimental design in providing improved taxonomic resolution.

Results: Extensive validation with simulated microbiome datasets mimicking different environmental and host-associated microbiome samples suggest that targeting certain combinations of non-contiguously placed V-regions might yield better taxonomic classification accuracy compared to conventional 16S amplicon sequencing targets. This work also puts forward a novel *in silico* combinatorial strategy that enables creation of consensus taxonomic profiles from experiments targeting multiple pair-wise combinations of V-regions to improve accuracy in taxonomic classification.

Conclusion: The study suggests that targeting non-contiguous V-regions with paired-end sequencing can improve 16S rRNA–based taxonomic resolution of microbiomes. Furthermore, employing the novel *in silico* combinatorial strategy can improve taxonomic

OPEN ACCESS

Edited by:
Jens Stoye,
Bielefeld University, Germany

Reviewed by:
Bernhard M. Fuchs,
Max Planck Institute for Marine Microbiology (MPG), Germany
Amnon Amir,
University of California, San Diego, United States

****Correspondence:***
Sharmila S. Mande
sharmila.mande@tcs.com
Anirban Dutta
anirban.dutt@tcs.com

[†]*These authors have contributed equally to this work.*

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

Received: *17 November 2018*
Accepted: *20 June 2019*
Published: *12 July 2019*

Citation:
Pinna NK, Dutta A, Monzoorul Haque M and Mande SS (2019) Can Targeting Non-Contiguous V-Regions With Paired-End Sequencing Improve 16S rRNA-Based Taxonomic Resolution of Microbiomes?: An In Silico Evaluation. Front. Genet. 10:653. doi: 10.3389/fgene.2019.00653

classification without any significant additional experimental costs and/or efforts. The empirical observations obtained can potentially serve as a guideline for future 16S microbiome studies, and facilitate researchers in choosing the optimal combination of V-regions for a specific experiment/sampled environment.

Keywords: metagenomics 16S, paired-end sequencing, taxonomic profiling, microbiome analysis, amplicon sequencing

INTRODUCTION

Sequencing of 16S rRNA genes is a standard protocol for taxonomic characterization of bacterial species (Schmalenberger et al., 2001; Clarridge, 2004; Munson et al., 2004; Petti et al., 2005). Sanger sequencing has been conventionally used for obtaining "full-length" 16S rRNA gene sequences of individual bacterium. Advent of next-generation sequencing (NGS) platforms has empowered the field of metagenomics and has enabled one to amplify and sequence (amplicon sequencing) specific portions of the 16S rRNA gene of community of bacteria (microbiome). Sequencing of such regions (encompassing one or more variable regions or V-regions) has been utilized in microbiome studies for obtaining taxonomic assignments for bacterial groups present in the studied environment. Although the accuracy and depth of taxonomic attribution obtained using such short reads are not at par as compared to that obtained using longer reads (Soergel et al., 2012; Martínez-Porchas et al., 2016), adoption of the former approach allows sequencing/sampling of large volumes of environmental DNA at significantly lower costs (Liu et al., 2012).

Depending on the sequencing platforms used, microbiome studies utilize either a single variable (V) region or a stretch of V-regions. For example, some of the Illumina platforms which generate very short reads (~150–250 base pairs in length) can be used to target only a single V-region using fragment library sequencing protocol (Bartram et al., 2011). On the other hand, technologies like Ion Torrent, Roche 454 etc., can generate longer reads (~400–500 bp) encompassing 2 or 3 contiguously placed V-regions (Loman et al., 2012; Salipante et al., 2014; D'Amore et al., 2016; Panek et al., 2018). Similar longer reads may also be targeted using a paired-end sequencing protocol on Illumina platforms (Fadrosh et al., 2014). It may also be noted that paired-end sequencing protocols, in principle, allows targeting and sequencing two sufficiently separated (non-contiguous) variable regions located on the same 16S rRNA gene (by choosing appropriate primers). Although paired-end sequencing has been in use for quite a while and have been used for whole-genome shotgun (WGS) sequencing-based metagenomics studies (Feng et al., 2015; Moustafa et al., 2018), to our knowledge, none of the 16S rRNA-based microbiome profiling studies have targeted or utilized a combination of "non-contiguous" V-regions for taxonomic characterization of bacterial communities. A few earlier studies have examined different aspects of short-read sequencing study designs with the goal of optimizing the choice of sequencing protocol (single-end *vs*. paired-end), target V-regions, as well as the taxonomic classification algorithm (Zhang et al., 2018; Yadav et al., 2019). A recent study has also attempted to combine taxonomic information from multiple V-regions (Fuks et al., 2018). Given the variable utility of different V-regions in resolving different bacterial taxonomic groups, it is also pertinent to ask whether the choice of V-regions should be restricted to a contiguous stretch, or be extended to a combination of V-regions placed "non-contiguously." To probe this at depth, we have performed comparison of taxonomic classifications obtained using various V-regions and their combinations. We have also assessed the feasibility of using "non-contiguous" V-region combinations for obtaining an accurate (and relatively higher resolution) taxonomic profile of a microbiome. The accuracy of taxonomic classifications obtained (at various levels of taxonomic hierarchy) using such non-contiguous V-regions has been compared with those obtained using single V-regions as well as with conventionally used combinations of contiguous V-regions.

METHODS

The primary objective of the current study involves evaluating/comparing the accuracy of taxonomic profiles generated from sequences encompassing (a) individual V-regions, (b) adjacent V-regions, and (c) pairs of non-contiguous V-regions and further assessing the discriminating capability of different V-regions with respect to different taxonomic lineages.

Full-length bacterial 16S rRNA gene sequences (along with their annotated lineages) present in the RDP database (release 11.3) (Cole et al., 2014) were downloaded for different analyses (described later in this section) in view of the abovementioned objectives. The RDP hierarchy browser (https://rdp.cme.msu.edu/hierarchy/hb_intro.jsp) was used for this purpose with the following filters—strain = "both"; source = "isolates"; size "> = 1,200"; taxonomy = "NCBI"; quality = "good," which resulted in a downloaded set of 232,163 sequences. Further, sequences not containing any of the nine V-regions (V1–V9) were filtered out from the set of sequences, leaving a total of 84,711 16S rRNA sequences belonging to 11,810 species, all of which contained all nine V-regions. Subsequently, both full-length as well as different portions of the 16S rRNA gene sequences were extracted *in silico* to represent outcomes of amplicon sequencing experiments and were provided as input to the Wang classifier (algorithm used in RDP classifier), as implemented in the software Mothur v.1.29.2 (Schloss et al., 2009), for taxonomic classification. The current version of RDP classifier 16S training set (https://sourceforge.net/projects/rdp-classifier/files/RDP_Classifier_TrainingData/RDPClassifier_16S_trainsetNo16_rawtrainingdata.zip/

download) was used as the reference database for these taxonomic assignment steps, and the taxonomic hierarchy information of the reference sequences were appropriately used while training the Wang classifier in order to enable obtaining taxonomic classifications resolved up to species level. Only a subset (57,632 sequences) of the originally downloaded full-length 16S rRNA gene sequences, which could be classified at species level with > = 80% bootstrap confidence threshold, was later used as a pool for randomly drawing sequences during creation of mock/simulated microbiome datasets (as described later in this section).

While evaluating the discriminating ability of individual V-regions, the regions of interest were parsed out from corresponding full-length 16S rRNA gene sequences using an in-house modified version of the V-Xtractor program (Hartmann et al., 2010), and submitted as query sequences to the Wang classifier. It may be noted in this context that reads generated during amplicon sequencing may often encompass flanking "constant" regions in addition to the targeted V-region(s), depending on choice of primers and the maximum read-length attainable by the sequencing technology. Consequently, our evaluation exercise, pertaining to combination of V-regions, aimed at mimicking 250 bp × 2 paired-end sequencing, wherein the extracted regions (representing sequenced reads) also encompass such flanking regions. To achieve this, regions from the full length 16S rRNA genes were extracted in such a way that either of the 250 bp reads (constituting a read-pair) contained one of the target V-regions, flanked in both directions by certain portions (lengths) of the surrounding "constant" regions. HMMs corresponding to constant regions surrounding the V-regions, as provided by the V-Xtractor program, were used for this purpose. Each extracted read started from a selected HMM near the target V-region (akin to a sequencing primer) and was extended to up to 250 bp toward the direction of the target V-region, thereby creating a read which encompassed the V-region along with some flanking sequence portion. It may be noted here that actual primer design may not always allow retention of flanks on either side of the targeted V-regions, equivalent to what was obtained using the HMMs, and results from an actual sequencing experiment may therefore slightly vary from the *in silico* validation results presented in this work. In case two adjacent V-regions were targeted, there was a significant chance of finding an overlap between two reads constituting a pair. This overlap was utilized to join the pair of reads together (used the program PEAR v0.9.6 with default parameters) (Zhang et al., 2014) into a single sequence before submitting the same as a query to the Wang classifier. In contrast, on targeting two distantly separated non-contiguous V-regions, no overlap between the read pairs could be expected. Accordingly, the pair of reads in this case were concatenated using a string of eight consecutive "*N*s," while preserving their orientation, prior to processing with Wang classifier. Given that Wang classifier (or RDP classifier) utilizes 8-mer nucleotide frequencies during taxonomic assignment (Wang et al., 2007), joining two non-overlapping sequenced fragments with 8 ambiguous nucleotides (N) ensures avoiding generation of spurious 8-mers consisting nucleotides from nonadjacent regions of the gene. The merging and concatenating of paired-end reads is depicted in a schematic diagram provided in **Supplementary Figure S1**. Taxonomic assignments generated by the Wang classifier at a predetermined taxonomic level with a confidence threshold score of > = 80% were used for all downstream comparative analyses. The different analyses performed and the underlying rationales are described in the following paragraphs.

First, the effectiveness of individual V-regions in resolving between different taxonomic groups was evaluated. For this purpose, different V-regions from all the 16S rRNA gene sequences, downloaded from the RDP database, were extracted. Subsequently, each of these individual V-regions were subjected to taxonomic classification with the Wang classifier (Wang et al., 2007), and the resultant assignments at the genus level were checked for accuracy and specificity against the taxonomic attributes provided by RDP for the corresponding full-length sequences.

The utility of all possible pair-wise combinations of V-regions, either arranged contiguously or non-contiguously, was also investigated *in silico* in terms of accuracy of taxonomic classifications provided by each such combination. As mentioned earlier, sequence fragments mimicking outcomes of 250 bp x 2 paired-end sequencing, which target different contiguous/non-contiguous combinations of V-regions, were derived from the downloaded 16S rRNA gene sequences. These fragments were subsequently subjected to taxonomic classification with the Wang classifier (Wang et al., 2007), and the assignments obtained at species level were checked for accuracy and specificity against the pre-annotated taxonomic attributes of their source (full-length) 16S rRNA genes.

The specific combinations of V-regions, which provided comparatively higher accuracies of taxonomic classification with the RDP database sequences, were further evaluated in a taxonomic assignment exercise with mock microbiome datasets. Five mock 16S microbiome gene pools were created from randomly selected sets of 50 organisms (genera) listed in RDP database (**Supplementary Table S1**). To obtain reads for building the mock microbiome datasets corresponding to these pools, each time, 10,000 16S rRNA genes were drawn randomly (following a uniform distribution) from a gene pool, such that the proportion of 16S rRNA genes drawn from any of the organisms are also randomized. Five such datasets (with 10,000 reads each) corresponding to each of the five gene pools (a total of 25 mock datasets) were constructed for comparative evaluation. Different contiguous as well as non-contiguous combinations of V-regions were subsequently extracted from each of the 16S rRNA genes belonging to these mock datasets and subjected to taxonomic analysis using Wang classifier, following the classification methodology described above. Taxonomic abundance values (obtained using different combinations of V-regions) were averaged over five mock datasets pertaining to the same gene pool. The averaged abundance values for each of the mock gene pools were compared against each other and the pre-annotated taxonomic attributes of their source (full-length) 16S rRNA genes, to assess the utility of the chosen combinations of V-regions. Nine more simulated microbiomes mimicking different environmental and host associated niches—namely, gut, skin, vaginal, sub-gingival (oral), sputum (oral), nematode gut, soil, and aquatic were also generated. Taxonomic abundance estimates for eight of these environmental microbiomes were derived from datasets used in an earlier *in silico* study evaluating functional potential of diverse metagenomes (Nagpal et al., 2016). Taxonomic abundance estimates for the aquatic microbiome was derived from a recent study by Muscarella and co-workers (Muscarella et al., 2019). To populate these simulated microbiomes, sequences from RDP database were randomly drawn (exact distributions provided in **Supplementary Table S2**), while making sure that the proportions of 16S rRNA genes drawn from different genera were roughly similar to the proportions observed earlier for these environments (Cui et al., 2012; Griffen et al., 2012; Human Microbiome Project Consortium, 2012; Alekseyenko et al., 2013; Botero et al., 2014; Kato et al., 2014; Romero et al., 2014; Xiao et al., 2014; Muscarella et al., 2019) (**Supplementary Table S3**). The taxonomic classification efficiency of the V-region combinations (at the species level) was also assessed on this set of simulated microbiomes.

In an ideal scenario, better taxonomic classification accuracy can be aimed for by using information from multiple V-regions. However, due to experimental limitations, this can be attained only if a long-read sequencing technology is used. To overcome this limitation, we propose a combinatorial strategy that extends the described paired-end sequencing workflow for targeting multiple pair-wise combinations of non-contiguous (or contiguous) V-regions in the following manner. The proposed strategy relies on obtaining taxonomic abundance profiles of a microbial community from two paired-end sequencing experiments, each of which targets different pair-wise combinations of V-regions. The two taxonomic profiles are then combined based on the accuracies of the individual V-regions (targeted in the experiments) in resolving each of the taxonomic groups under consideration. **Figure 1** and the following generic example illustrate the strategy in detail: A

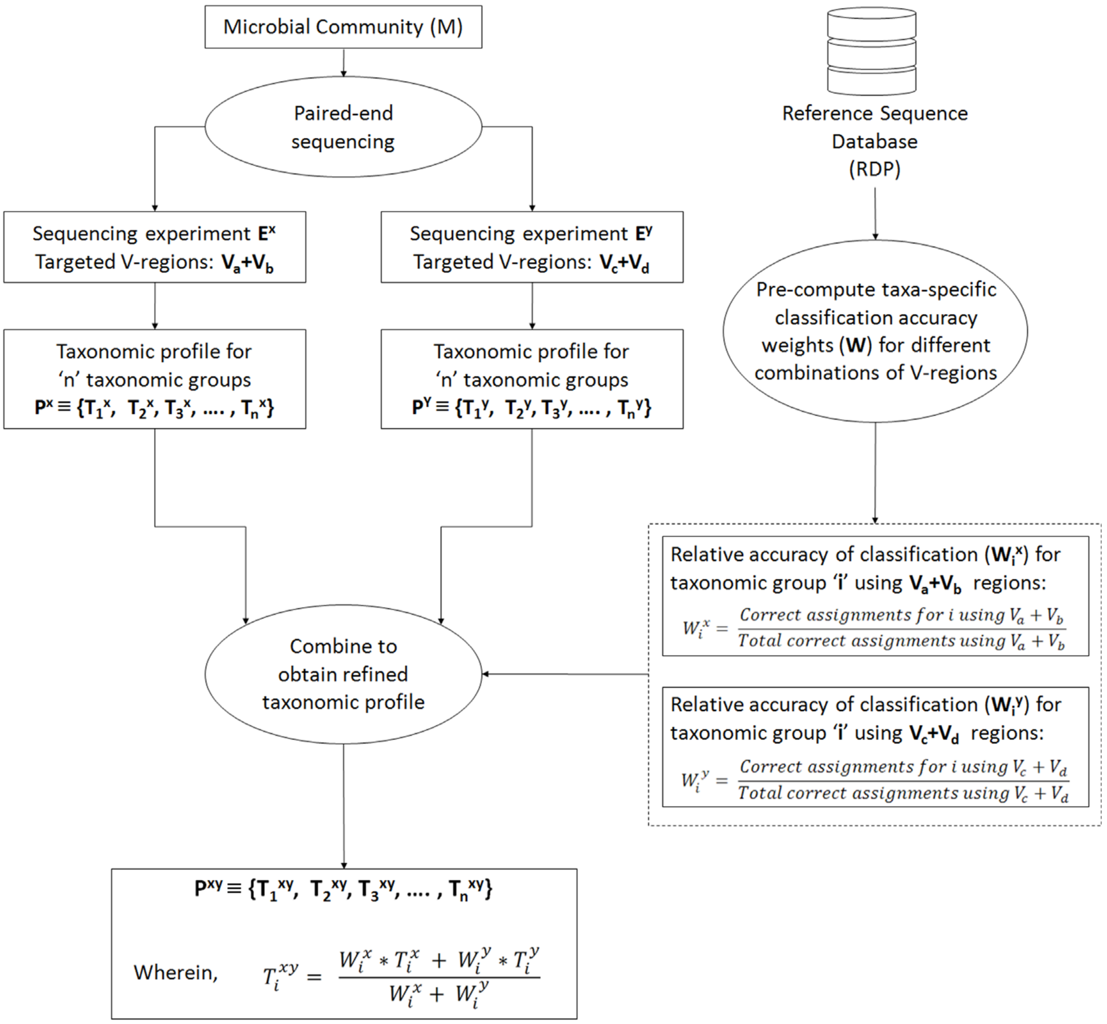

Microbial Community (M)
Paired-end sequencing
Reference Sequence Database (RDP)
Sequencing experiment $\mathbf{E^x}$
Targeted V-regions: $\mathbf{V_a+V_b}$
Sequencing experiment $\mathbf{E^y}$
Targeted V-regions: $\mathbf{V_c+V_d}$
Pre-compute taxa-specific classification accuracy weights (**W**) for different combinations of V-regions
Taxonomic profile for 'n' taxonomic groups
$\mathbf{P^x \equiv \{T_1^x,\ T_2^x, T_3^x, \ldots, T_n^x\}}$
Taxonomic profile for 'n' taxonomic groups
$\mathbf{P^Y \equiv \{T_1^y,\ T_2^y, T_3^y, \ldots, T_n^y\}}$
Relative accuracy of classification ($\mathbf{W_i^x}$) for taxonomic group **'i'** using $\mathbf{V_a+V_b}$ regions:
$W_i^x = \frac{Correct\ assignments\ for\ i\ using\ V_a + V_b}{Total\ correct\ assignments\ using\ V_a + V_b}$
Combine to obtain refined taxonomic profile
Relative accuracy of classification ($\mathbf{W_i^y}$) for taxonomic group **'i'** using $\mathbf{V_c+V_d}$ regions:
$W_i^y = \frac{Correct\ assignments\ for\ i\ using\ V_c + V_d}{Total\ correct\ assignments\ using\ V_c + V_d}$
$\mathbf{P^{xy} \equiv \{T_1^{xy},\ T_2^{xy}, T_3^{xy}, \ldots, T_n^{xy}\}}$
Wherein, $T_i^{xy} = \frac{W_i^x * T_i^x + W_i^y * T_i^y}{W_i^x + W_i^y}$

FIGURE 1 | Combinatorial strategy for targeting multiple pair-wise combinations of non-contiguous (or contiguous) V-regions. The strategy relies on obtaining taxonomic abundance profiles of a microbial community from two paired-end sequencing experiments, each of which targets different pair-wise combinations of V-regions. The two taxonomic profiles are then combined based on the pre-calculated accuracies of individual V-regions (targeted in the two experiments) in resolving each of the taxonomic groups under consideration.

microbial community (M) is initially considered for taxonomic profiling by two paired-end sequencing experiments (E^x and E^y). Each of these experiments can target two distinct V-regions (either arranged contiguously or non-contiguously on the 16S rRNA gene), using appropriate forward and reverse primers, as described in the previous sections. Let us consider that in the current example, E^x targets the V-region combination V_a+V_b, and E^y targets V_c+V_d. For example, combinations of V-regions selected in the two experiments could be V1+V4 and V2+V6 in one scenario. Based on the taxonomic resolution efficiencies of different (combinations of) V-regions, E_x and E_y will generate two different taxonomic abundance profiles P^x and P^y, respectively, each of which constitutes of estimated abundance values (T_i) for different taxonomic groups (i):

$$P^x \equiv \{T_1^x, T_2^x, T_3^x, \ldots\ldots, T_n^x\} \qquad \text{Equation 1}$$

$$P^x \equiv \{T_1^y, T_2^y, T_3^y, \ldots\ldots, T_n^y\} \qquad \text{Equation 2}$$

Subsequently, for each of the taxonomic groups (T_i), a refined estimate of its abundance (T_i^{xy}) can be arrived at by combining the observed abundances T_i^x and T_i^y, such that the refined abundance T_i^{xy} is relatively closer to the estimate obtained with the experiment (either of E^x or E^y) providing better classification accuracies for taxa 'i.' Calculation of the refined estimate therefore takes into consideration the taxonomic classification accuracies of the combination of V-regions that had been used for the initial set of experiments E^x and E^y using the following equation:

$$T_i^{xy} = \frac{W_i^x * T_i^x + W_i^y * T_i^y}{W_i^x + W_i^y} \qquad \text{Equation 3}$$

wherein W_i^x and W_i^y are the relative accuracies in taxonomic classification for a particular taxonomic group 'i,' obtained using the specific combination of V-regions chosen for experiments E^x and E^y respectively. In case the refined taxonomic profiles are to be represented in terms of normalized abundance values, e.g., frequencies or percentage normalized abundances, the refined T_i^{xy} values from equation 3 needs to be appropriately modified (normalized) further. This weighted average approach has been adopted considering that different V-regions (or their combinations) have different efficiencies in resolving the same taxonomic group. A simple average therefore would not be appropriate for combining two taxonomic abundance estimates pertaining to a sample, which has been generated through separate experiments targeting different V-regions (or their combinations). Instead, the refined taxonomic abundance value for a given taxon should be weighted toward the results generated by the V-region (or a combination) which is more accurate in classifying the taxon in question. These accuracies can be calculated from the evaluation results obtained from **Supplementary Table S4**, as a ratio of the correct assignments obtained for particular taxa using a specific combination of V-regions, and the total number of correct assignments obtained using the same V-region combination. For example, considering that the combination of V_a+V_b was used in experiment E^x, W_i^x can be calculated as

$$W_i^x = \frac{\textit{Correct assignments for taxon i using } V_a + V_b}{\textit{Total correct assignments using } V_a + V_b} \qquad \text{Equation 4}$$

The denominator term representing "total correct assignments using V_a+V_b" has been introduced to capture any additional specificity of the chosen V_a+V_b region toward a particular taxon 'i' in context of the overall taxonomic classification performance of V_a+V_b. Other simple ways of calculating the "relative accuracy in taxonomic classification" or weight (W_i^x), e.g., in a case wherein the denominator term is omitted, would also work fine when V-region combinations with decent classification accuracy are chosen. It may be noted here that in the experiment(s) using paired-end sequencing to capture two different V-regions from the 16S rRNA gene, the correspondence between the pairs of V-regions originating from the same 16S rRNA gene is retained. This allows joining the different V-regions together into a single DNA string (separated appropriately by ambiguous nucleotide characters) and providing the same as an input to taxonomic classification tools, such as the RDP classifier. However, for V-regions targeted in separate sequencing experiments, cross-experiment correspondence between the sequenced V-regions with respect to their origin 16S rRNA gene cannot be identified. This necessitates the indirect strategy of combining information obtained from different V-regions (or their combinations) for refining the taxonomic abundance estimates, as described above. To avoid variations arising from experimental workflows and sample handling/preparations, it would be ideal to perform a single PCR step for amplicon generation, using different sets of primers appropriate for the chosen combinations of V-regions (V_a+V_b, and V_c+V_d in the given example). However, it also needs to be mentioned here that the designed primers may have different affinities for the targeted regions on 16S rRNA genes originating from different taxonomic groups. This may again result in unequal proportions of 16S rRNA sequence fragments amplified by the different sets of primers, which would subsequently be reflected in the sequencing outcome. In such a scenario, the combination strategy needs to factor in this difference in proportions, while arriving at a refined taxonomic abundance estimate. Alternately, the experiment may target a combination of 3 V-regions (e.g., V_a+V_b and V_a+V_c or, V_a+V_c and V_b+V_c), such that either the forward primers or the reverse primers be common to the targeted combinations. This way, some equivalence in the proportions of fragments (targeting different taxonomic groups) can be maintained on account of the shared primer (for V-region) selected.

To asses the utility of the combinatorial strategy, the taxonomic abundance profile of the simulated microbiome sample pertaining to human gut (as described earlier) was re-evaluated, targeting the V-region combinations V_1+V_4 and V_1+V_5, both of which had decent classification accuracies. 5,000 sequence fragments corresponding to each of the V-region combinations (i.e., a total of 10,000 fragments) were sampled from the simulated gut microbiome. The results obtained with the combinatorial strategy was subsequently compared against the results obtained when each of the V-region combinations were used separately. To maintain equivalence in sequencing coverage, 10,000 fragments were sampled from the simulated gut microbiome, while targeting the V-region combinations separately.

RESULTS AND DISCUSSION

Individual V-Regions Have Differential Ability in Resolving Various Taxonomic Groups

The accuracies of different V-regions in resolving different taxonomic groups are depicted in **Figure 2**. The classification accuracies (at genus level) obtained with V-regions have been cumulated and depicted at the "phylum level" in the figure and placed in context with the classification accuracies which would have been obtained with full-length 16S rRNA gene sequences (details in Methods). Except for V1, V5, and V9, all other V-regions were observed to have certain utility in taxonomic classification, even when targeted individually. It was also evident from the plot that some V-regions provide comparatively higher accuracies of classification for specific taxonomic groups. For example, the V4 region has the highest accuracy while classifying sequences pertaining to the phylum *Bacteroidetes* (75.9%), whereas the V2 region classifies best with respect to the phylum *Firmicutes* (68.2%). However, it may be noted that a sequenced read generated in a real amplicon sequencing experiment will extend beyond the targeted V-regions and include some surrounding portions. The resultant taxonomic classification in such a case is expected to be better than the currently depicted results which were generated based on the exact V-regions. A detailed list of accuracies in taxonomic classification obtained with different V-regions at genus level is provided in **Supplementary Table S5**. Given these observations, it would seem logical for a microbiome study design to sequence two (or more) V-regions from a 16S rRNA gene fragment which have complementary abilities with respect to classification of different taxonomic groups. Furthermore, the choice of the combination of V-regions could also be guided by the environment from where the microbiome sample is being collected, given that diverse environments may be differentially enriched with different taxonomic groups.

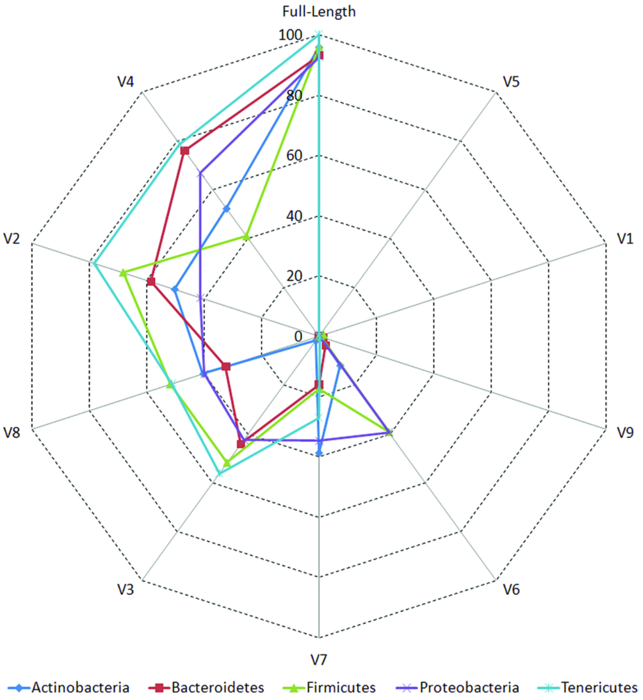


FIGURE 2 | Taxonomic classification accuracies at genus level for different variable regions. Plot depicting the percentage of 16S rRNA genes present in RDP database that could be correctly classified utilizing different variable (V) regions (see Methods). Correct classifications obtained using full-length 16S sequences are also depicted for comparison. Taxonomic classification accuracy at genus level has been considered in this plot and has been cumulated and depicted at the phylum level (only for five most represented phyla in the downloaded RDP sequences).

A preferred combination of V-regions cannot always be expected to be situated in a contiguous stretch on the 16S rRNA gene. Given the read length limitations of NGS technologies, targeting an amplicon constituting the preferred regions becomes difficult in reality. The length distributions of V-regions and C-regions (constant/conserved regions flanking the V-regions) across different bacterial taxonomic groups are provided in **Supplementary Figure S2**. These distributions indicate that while individual V-regions and contiguous stretches like V2–V3 (median length 297 bp) or V3–V4 (median length 254 bp) can easily be targeted with short-read sequencing techniques like Illumina HiSeq/MiSeq, sequencing longer contiguous stretches encompassing more than two V-regions, such as V2-V3-V4 (median length 482 bp) and V4-V5-V6 (median length 453 bp), necessitates sequencing platforms that can generate longer read lengths (e.g., Roche 454). Capturing even more V-regions on a single read is beyond the scope of most current generation high-throughput sequencing technologies. Consequently, targeting an optimal combination of V-regions, which may be present on the 16S rRNA gene in either contiguous or non-contiguous arrangement(s), remains a challenge.

Targeting Combinations of Non-Contiguously Placed V-Regions Using Paired-End Sequencing Enables Improved Taxonomic Classification

Paired-end sequencing protocols available with some of the NGS platforms allow sequencing of a stretch of DNA from both its ends (Rodrigue et al., 2010; Dutta et al., 2014). For example, Illumina HiSeq sequencing platforms can be used for paired-end sequencing to generate up to 2x250bp reads. The current work proposes, and evaluates *in silico*, the utilization of paired-end sequencing protocols for sequencing various pair-wise combinations of non-contiguous V-regions in a single sequencing run. To this end, appropriate primers need to be designed against a desired stretch of the 16S rRNA gene, such that the targeted V-regions (either contiguously or non-contiguously placed) reside within this stretch and are not far from either of its boundaries. Sequencing of the amplicons generated with these primers can then be performed with a paired-end sequencing protocol, whereby these (amplified) stretches of DNA are sequenced from both ends. Two reads sequenced from each such amplicon would cover the two targeted V-regions (one from each end). Since each of the sequenced reads from any given "pair" targets a single V-region (situated at one of the ends of the amplicon), read-length limitations do not restrict capturing the entirety of the individual V-regions. Consequently, it becomes possible to sequence almost all possible pair-wise combinations of V-regions, either arranged contiguously or non-contiguously.

The results pertaining to the *in silico* evaluation of the effectiveness of different combinations of V-regions (see Methods), in providing accurate taxonomic classifications (at the species level) for sequences listed in the RDP database, is depicted in **Figure 3** (also see **Supplementary Table S4**).

Classification accuracies provided by several combinations of non-contiguously placed V-region pairs, namely, V1+V3 (77.7%), V1+V4 (77.4%), V1+V8 (76.6%), V2+V5 (73.6%), etc., were sufficiently high and exceeded the classification accuracies provided by even the best of the combinations of adjacently placed V-regions (e.g., 68.6% by V1+V2, 70.9% by V2+V3) by a fair margin of 5–8%. It was also significant to note that many of the individual V-regions, which had very low taxonomic discriminating ability of their own (**Figure 2**, **Supplementary Table S5**), could provide significant classification accuracies when paired up with other V-regions. For example, while V1 and V5 provided very low taxonomic classification accuracies when targeted alone, the combination of V1+V5 could provide a significantly high taxonomic classification accuracy of 73.4%. Furthermore, although the individual V-regions were observed to have differential abilities in classifying sequences originating from different phyla (**Figure 2**), their combinations were much more coherent in this regard and could classify sequences from all phyla with better efficiency (**Figure 4**) than single V-regions. Results indicate the potential utility of targeting pairs of non-contiguously placed V-regions to improve taxonomic classification accuracy. Additionally, the results also suggest that for exploring the taxonomic diversity of a particular environment, which may be expected to be enriched with particular groups of bacteria, an appropriate combination of V-regions sensitive to the same bacterial groups may be chosen.

To assess the utility of the proposed non-contiguous combination of V-regions on a microbiome dataset, while avoiding any bias arising out of the proportion of sequences pertaining to different bacterial groups currently catalogued in reference databases like RDP, taxonomic classification exercises were further performed with mock microbiome datasets. Each of the mock microbiome datasets were constructed using 10,000 randomly selected 16S rRNA gene sequences from one of the five randomized 16S gene pools. Each of these gene pools consisted of sequences downloaded from the RDP database, wherein the proportion of sequences selected from different organisms were also randomized (see Methods). The results, in terms of classification accuracy at the species level, are depicted in **Table 1**. It was interesting to note that 18 out of the 20 combinations of V-regions, which could provide classification accuracy > = 60% on average, constituted of non-contiguous V-regions. The best performing combination of adjacent V-regions was V2–V3, which on average provided 69.1% classification accuracy. In comparison, the combination of the non-contiguously placed V-regions V1+V4 demonstrated a high average classification accuracy of 77.2%.

The efficiency of the proposed non-contiguous combination of V-regions was further tested on nine additional simulated microbiomes (see Methods) mimicking different environmental and host-associated niches (see Methods, **Supplementary Tables S3**, **S2**, and **S6**). Results pertaining to these simulated microbiomes—namely, gut, skin, vaginal, sub-gingival (oral), sputum (oral), nematode gut, soil, and aquatic are depicted in **Figure 5**. It was interesting to note that optimal classification of reads from the simulated microbiomes pertaining to different

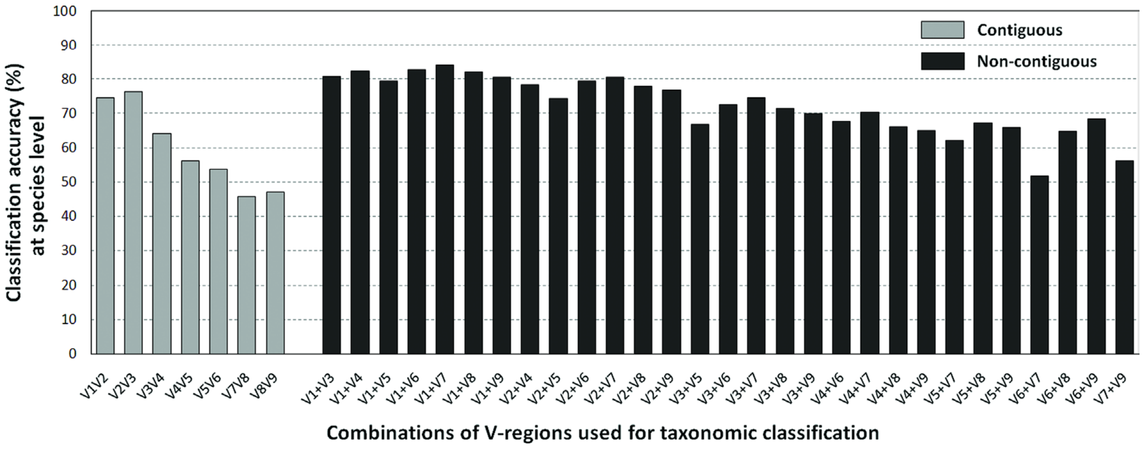


FIGURE 3 | Taxonomic classification accuracies at species level for different variable regions. Plot depicting the average taxonomic classification accuracies obtained at species level using different pair-wise combinations of V-regions (both contiguous as well as non-contiguous) drawn from the 16S rRNA genes. 16S rRNA genes used for the evaluation were retrieved from the RDP database (see Methods).

niches could be obtained with different combinations of non-contiguous V-regions.

The combination of V1+V4 regions provided the maximum accuracy of classification for skin (60.2%) and one of the gut (86.0%) microbiomes (Gut2), whereas microbiomes pertaining to vaginal and sub-gingival niches were best resolved by the combination V1+V9 (with accuracies of 83.3% and 78.6%, respectively). Optimal classification of sputum microbiome samples (72.1%) could be obtained by another non-contiguous combination, viz. V1+V5 regions, which could also provide relatively more accurate classification for the Gut1 microbiome (82.5%). It was also interesting to note the high variability in classification accuracies of individual V-region combinations while classifying samples pertaining to different environments. For example, while the combination V2+V4 could classify one of the gut microbiomes (Gut2) with 85.93% accuracy, the classification results were not as high when the same combination was used to classify the aquatic microbiome (69.2%). On the other hand, the combination V2+V7 was observed to provide decent classification for the simulated aquatic microbiome (72.8%), while performing not so well for the simulated gut microbiome datasets (65.8% for Gut1 and 70.9% for Gut2). These results further reiterate the need of choosing an optimal combination of V-regions, preferably non-contiguous, for a specific sampled environment.

It may be noted here that the paired-end reads generated for *in silico* evaluation of the utility of different combinations of V-regions were based on HMMs pertaining to the flanking constant regions, as provided by the V-Xtractor program (see Methods). Actual primer design may not always allow generation of reads identical to the *in silico* experiment, and results from a sequencing experiment may slightly vary from the validation results presented. A comparison of the paired-end reads generated in the *in silico* experiments with respect to those which may be obtained by using different sets of primers currently available for 16S rRNA amplicon sequencing is provided in **Supplementary Figure S3**, and **Supplementary Tables S7** and **S8**. **Supplementary Figure S3(A)** and **Supplementary Table S8** additionally depict

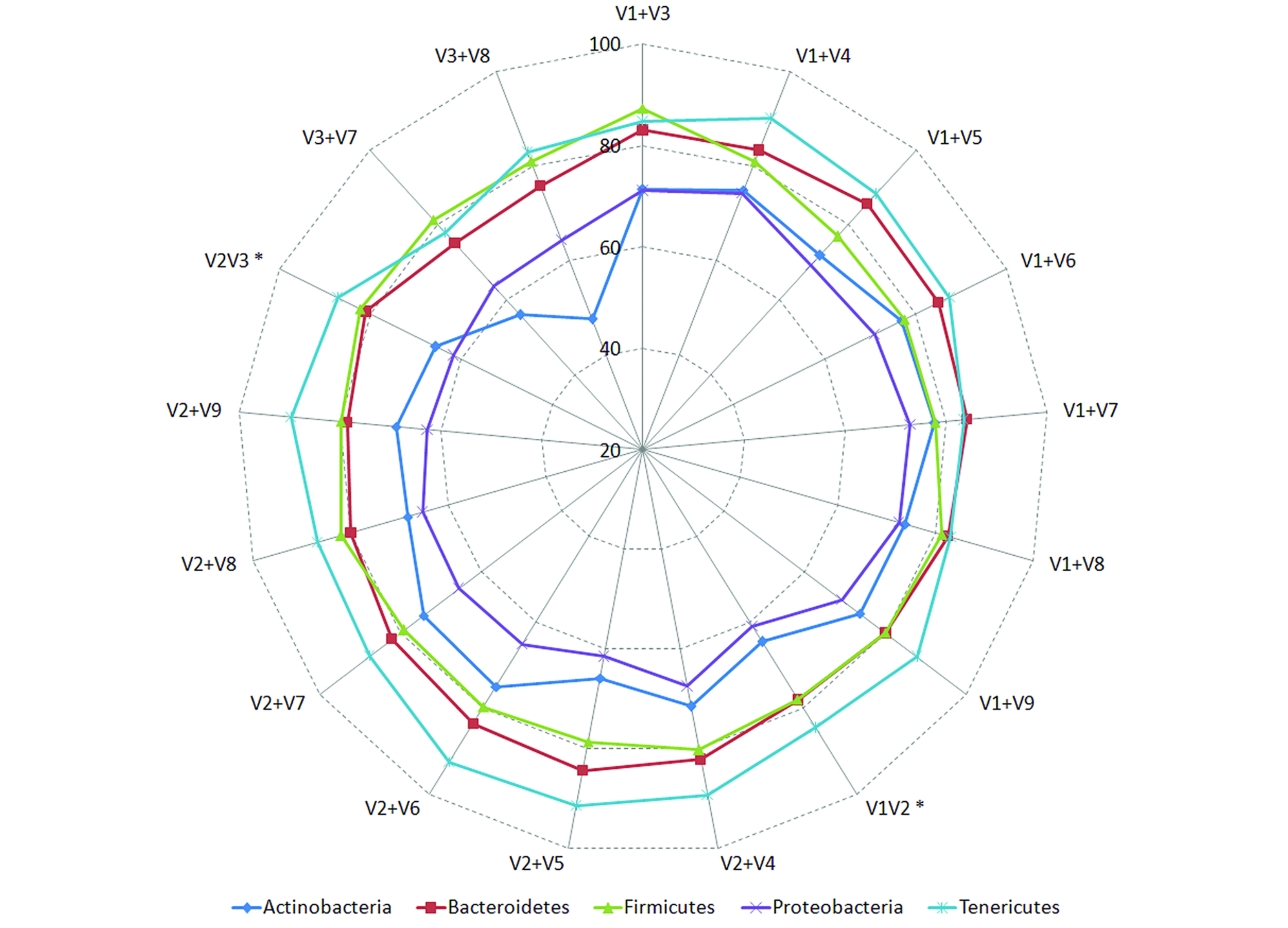


FIGURE 4 | Taxonomic classification accuracies obtained using different pair-wise combinations of V-regions (contiguous as well as non-contiguous). Accuracy of taxonomic assignments has been evaluated at the species level and cumulated at phylum level for representation (only for five most represented phyla in the downloaded RDP sequences). Combinations of V-regions achieving a classification accuracy of > = 70% (averaged for the depicted phyla) are shown. Combinations of contiguously placed V-regions have been indicated with an asterisk (*).

TABLE 1 | Taxonomic classification accuracies obtained using different pair-wise combinations of V-regions (both contiguous as well as non-contiguous) evaluated for mock microbiome datasets, each constituting of 10,000 randomly selected 16S rRNA genes from five different 16S gene pools.

Combination of V-region	Classification accuracy (%) at species level averaged over five mock datasets from each 16S gene pool					
	Mock datasets from 16S gene pool 1	Mock datasets from 16S gene pool 2	Mock datasets from 16S gene pool 3	Mock datasets from 16S gene pool 4	Mock datasets from 16S gene pool 5	Average accuracy
V1+V4	77.29	79.47	72.79	75.90	80.48	77.19
V1+V3	74.69	78.16	77.52	74.76	80.08	77.04
V1+V8	76.03	77.96	73.24	75.72	79.32	76.46
V1+V7	77.20	78.33	70.37	77.34	78.60	76.37
V1+V6	72.46	77.34	69.73	78.25	76.90	74.94
V1+V5	70.89	74.24	69.16	73.37	76.40	72.81
V1+V9	71.74	71.41	71.33	73.95	75.57	72.80
V2+V4	69.07	75.07	72.76	70.99	73.55	72.29
V2+V8	68.26	74.60	73.33	70.66	73.27	72.02
V2+V6	66.84	74.54	72.60	72.19	72.67	71.77
V2+V7	68.34	72.76	72.73	71.17	71.30	71.26
V2V3*	61.53	71.52	72.03	66.31	73.92	69.06
V2+V9	65.03	68.85	71.60	66.32	71.81	68.72
V1V2*	64.20	70.29	66.81	65.44	72.40	67.83
V3+V8	68.47	61.80	69.66	66.59	67.82	66.87
V3+V7	68.41	61.60	71.05	66.80	65.93	66.76
V2+V5	61.38	68.19	68.42	65.36	69.34	66.54
V3+V6	63.26	59.91	68.53	67.04	65.15	64.78
V3+V9	63.63	55.85	67.20	65.94	63.83	63.29
V3+V5	60.94	56.74	65.79	62.91	62.49	61.77

Accuracy of taxonomic assignments has been evaluated at the species level considering the assignments obtained with full-length 16S sequences to be correct. Top 20 combinations in terms of average classification accuracy have been depicted. Combinations of contiguous V-regions have been marked with an asterisk (*).

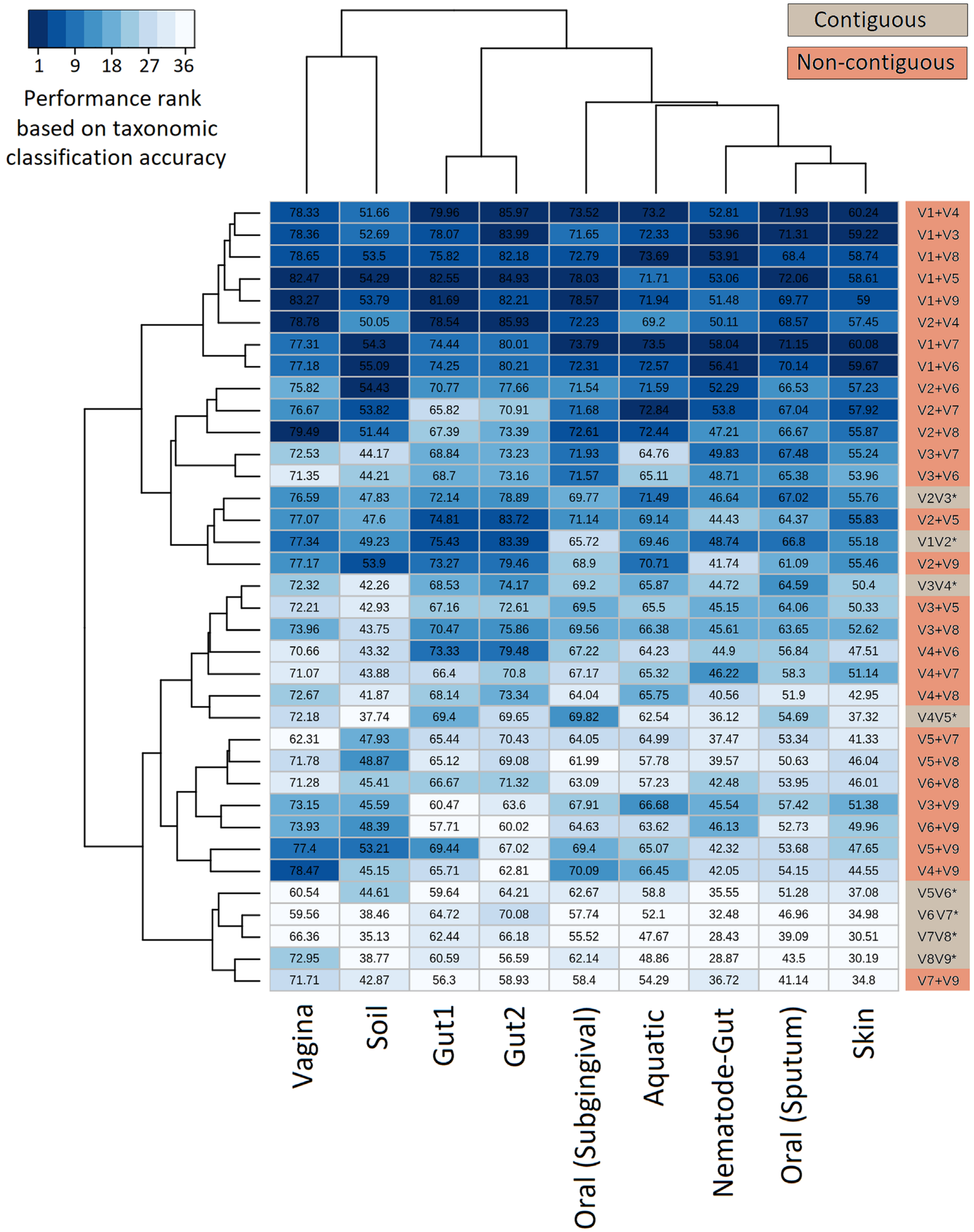
the specificity of different primer sets that may be used to target various combinations of V-regions with respect to the sequences present in the RDP database. It may be mentioned here that assessment of primer specificity on all sequences from RDP database (a total of 232,163 sequences having length > = 1,200 bp) revealed that the combinations/pairs (either contiguous or non contiguous) involving the V1-region could potentially amplify a lower fraction of sequences compared to other combinations. Apparently, the fraction of sequences that can be amplified by the said combinations is limited by the specificity/universality of the primer for V1-region. The presence of many incomplete/truncated SSU rRNA sequences in RDP database, which might be missing the V1 primer binding sites may also contribute to this observation. The overall results, however, do not indicate any significant deviations in the specificity (fraction of bacterial sequences amplified) of primer pairs targeting non-contiguous V-regions, when compared to the primers targeting contiguously placed V-regions.

It may also be noted that this work did not compare and validate the efficacy of the proposed method in perspective of some recent taxonomic analysis methods which performs exact sequence variant (ESV) analyses (Amir et al., 2017; Callahan et al., 2016, 2). This was primarily because the currently available implementations of such methods expect a significant overlap between the paired-end reads and only work after the two reads are merged (or works with individual reads), thereby making it difficult to make a direct comparison with non-overlapping paired-end reads targeting non-contiguous V-regions. However, one would expect that the combination of V-regions that cannot provide good resolution at genus or species levels will also fail at deeper taxonomic levels like OTUs/sub-OTUs/ESVs, and vice versa.

Consensus of Multiple Combinations of V-Regions Enables Further Refinement of Taxonomic Profiles

Although better taxonomic classification accuracies can be obtained by using information from multiple V-regions, relatively higher costs and lower throughput serve as deterrents against adoption of long-read sequencing technologies for metagenomic studies. To overcome this bottleneck, we propose a combinatorial strategy (**Figure 1**) that extends the described paired-end sequencing workflow (achievable with a short read sequencing technology like Illumina) for targeting multiple pair-wise combinations of non-contiguous (or contiguous) V-regions (see Methods). The proposed strategy relies on obtaining taxonomic abundance profiles of a microbial community from two paired-end sequencing experiments, each of which targets different pair-wise combinations of V-regions. The two taxonomic profiles are then combined based on (pre-estimated) accuracies of the individual V-regions (targeted in the experiments) in resolving each of the taxonomic groups under consideration.

Considering the fact that human gut is one of the most diverse and densely populated reservoir of microbes, the utility of the combinatorial strategy was assessed with one of the simulated human gut microbiome sample Gut1 (as described earlier). As can be seen from **Figure 5**, the V-region combinations V$_1$+V$_4$



	Vagina	Soil	Gut1	Gut2	Oral (Subgingival)	Aquatic	Nematode-Gut	Oral (Sputum)	Skin
V1+V4	78.33	51.66	79.96	85.97	73.52	73.2	52.81	71.93	60.24
V1+V3	78.36	52.69	78.07	83.99	71.65	72.33	53.96	71.31	59.22
V1+V8	78.65	53.5	75.82	82.18	72.79	73.69	53.91	68.4	58.74
V1+V5	82.47	54.29	82.55	84.93	78.03	71.71	53.06	72.06	58.61
V1+V9	83.27	53.79	81.69	82.21	78.57	71.94	51.48	69.77	59
V2+V4	78.78	50.05	78.54	85.93	72.23	69.2	50.11	68.57	57.45
V1+V7	77.31	54.3	74.44	80.01	73.79	73.5	58.04	71.15	60.08
V1+V6	77.18	55.09	74.25	80.21	72.31	72.57	56.41	70.14	59.67
V2+V6	75.82	54.43	70.77	77.66	71.54	71.59	52.29	66.53	57.23
V2+V7	76.67	53.82	65.82	70.91	71.68	72.84	53.8	67.04	57.92
V2+V8	79.49	51.44	67.39	73.39	72.61	72.44	47.21	66.67	55.87
V3+V7	72.53	44.17	68.84	73.23	71.93	64.76	49.83	67.48	55.24
V3+V6	71.35	44.21	68.7	73.16	71.57	65.11	48.71	65.38	53.96
V2V3*	76.59	47.83	72.14	78.89	69.77	71.49	46.64	67.02	55.76
V2+V5	77.07	47.6	74.81	83.72	71.14	69.14	44.43	64.37	55.83
V1V2*	77.34	49.23	75.43	83.39	65.72	69.46	48.74	66.8	55.18
V2+V9	77.17	53.9	73.27	79.46	68.9	70.71	41.74	61.09	55.46
V3V4*	72.32	42.26	68.53	74.17	69.2	65.87	44.72	64.59	50.4
V3+V5	72.21	42.93	67.16	72.61	69.5	65.5	45.15	64.06	50.33
V3+V8	73.96	43.75	70.47	75.86	69.56	66.38	45.61	63.65	52.62
V4+V6	70.66	43.32	73.33	79.48	67.22	64.23	44.9	56.84	47.51
V4+V7	71.07	43.88	66.4	70.8	67.17	65.32	46.22	58.3	51.14
V4+V8	72.67	41.87	68.14	73.34	64.04	65.75	40.56	51.9	42.95
V4V5*	72.18	37.74	69.4	69.65	69.82	62.54	36.12	54.69	37.32
V5+V7	62.31	47.93	65.44	70.43	64.05	64.99	37.47	53.34	41.33
V5+V8	71.78	48.87	65.12	69.08	61.99	57.78	39.57	50.63	46.04
V6+V8	71.28	45.41	66.67	71.32	63.09	57.23	42.48	53.95	46.01
V3+V9	73.15	45.59	60.47	63.6	67.91	66.68	45.54	57.42	51.38
V6+V9	73.93	48.39	57.71	60.02	64.63	63.62	46.13	52.73	49.96
V5+V9	77.4	53.21	69.44	67.02	69.4	65.07	42.32	53.68	47.65
V4+V9	78.47	45.15	65.71	62.81	70.09	66.45	42.05	54.15	44.55
V5V6*	60.54	44.61	59.64	64.21	62.67	58.8	35.55	51.28	37.08
V6V7*	59.56	38.46	64.72	70.08	57.74	52.1	32.48	46.96	34.98
V7V8*	66.36	35.13	62.44	66.18	55.52	47.67	28.43	39.09	30.51
V8V9*	72.95	38.77	60.59	56.59	62.14	48.86	28.87	43.5	30.19
V7+V9	71.71	42.87	56.3	58.93	58.4	54.29	36.72	41.14	34.8

FIGURE 5 | Evaluation of taxonomic classification efficiency on simulated microbiomes. Taxonomic classification efficiency of different combinations of V-regions evaluated on nine simulated microbiome datasets mimicking different environmental niches. Taxonomic classification accuracy in terms of percentages of correct assignments at species level are indicated in the heatmap. The color scale (1–36) depicts the performance rank of different combinations of V-regions (total of 36 combinations) in terms of taxonomic classification accuracy for each of the simulated microbiomes (presented in columns).

TABLE 2 | Utility of proposed combinatorial approach in obtaining refined taxonomic profiles compared to taxonomic abundance estimates obtained with pair-wise combinations of V-regions.

Species	Abundance (%) estimated with full-length 16S reads	Abundance (%) estimated with 10,000 V1+V4 paired-end reads	Abundance (%) estimated with 10,000 V1+V5 paired-end reads	Abundance (%) estimated with combinatorial approach using 5,000 V1+V4 and 5,000 V1+V5 reads
Faecalibacterium prausnitzii	11.17	12.24	12.25	11.06
Bacteroides faecis	10.69	11.97	11.24	11.36
Prevotella amnii	6.73	0.00	6.72	7.28
Prevotella nigrescens	6.47	6.98	6.76	6.96
Megamonas hypermegale	5.35	6.06	3.53	4.71
Bacteroides pyogenes	4.23	4.44	4.33	4.55
Bacteroides finegoldii	3.98	4.03	4.13	4.00
Alistipes putredinis	3.45	3.73	3.71	3.51
Roseburia hominis	2.41	2.70	2.84	2.62
Bacteroides nordii	2.18	2.50	2.26	2.16
Bacteroides eggerthii	2.15	2.51	2.24	2.15
Bacteroides helcogenes	2.09	2.35	2.13	2.11
Bacteroides caccae	2.08	2.30	2.32	2.32
Bacteroides massiliensis	2.07	2.10	2.13	2.03
Bacteroides coprocola	2.04	2.43	2.27	2.21
Bacteroides salyersiae	2.04	2.26	2.01	2.12
Bacteroides stercoris	2.03	1.92	2.50	2.17
Bacteroides uniformis	2.02	2.03	2.04	1.93
Bacteroides acidifaciens	2.01	2.30	2.00	2.08
Proteiniphilum acetatigenes	2.01	2.21	2.18	2.07
Bacteroides cellulosilyticus	1.98	2.16	0.00	1.70
Bacteroides intestinalis	1.96	2.02	2.08	2.03
Roseburia faecis	1.74	1.94	1.91	1.69
Roseburia intestinalis	1.74	2.16	1.91	1.86
Parasutterella secunda	1.50	1.74	1.56	1.38
Roseburia inulinivorans	1.00	1.00	1.06	1.11
Phascolarctobacterium succinatutens YIT 12067	0.99	0.82	0.78	0.80
Parabacteroides distasonis	0.90	1.03	1.04	0.74
Parabacteroides merdae	0.89	1.07	0.87	0.92
Parasutterella excrementihominis	0.82	0.99	0.84	0.75
Dorea longicatena	0.78	0.32	0.51	0.32
Phascolarctobacterium faecium	0.74	0.81	0.83	0.69
Blautia producta	0.70	0.55	0.86	0.61
Escherichia/Shigella fergusonii	0.69	0.59	0.00	0.00
Escherichia/Shigella albertii	0.57	0.56	0.62	0.71
Escherichia/Shigella flexneri	0.56	0.00	0.00	0.00
Escherichia/Shigella dysenteriae	0.53	0.50	0.54	0.57
Dialister invisus	0.47	0.58	0.50	0.46
Megasphaera elsdenii	0.46	0.37	0.47	0.40
Blautiaglucerasea	0.45	0.41	0.48	0.61
Blautia hydrogenotrophica	0.43	0.44	0.46	0.51
Blautia schinkii	0.43	0.47	0.54	0.43
Mitsuokella jalaludinii	0.39	0.40	0.42	0.35
Collinsella aerofaciens	0.34	0.37	0.42	0.36
Bifidobacterium longum	0.32	0.40	0.37	0.38
Bifidobacterium animalis	0.32	0.25	0.32	0.29
Ruminococcus flavefaciens	0.30	0.21	0.25	0.17
Blautia hansenii	0.28	0.33	0.30	0.33
Megasphaera sp. *NMBHI-10*	0.28	0.22	0.19	0.17
Klebsiella pneumoniae	0.25	0.21	0.29	0.27
Cumulated percentage deviation from abundance estimated using full-length 16S sequences	**–**	**17.40**	**11.47**	**6.85**

Results in the table pertain to the simulated human gut microbiome dataset Gut1 (as depicted in **Figure 5***).*

and V_1+V_5 provided highest average classification accuracies for most of the host (human)-associated environmental niches. Consequently, these V-region combinations were targeted for evaluating this combinatorial strategy wherein 5,000 sequence fragments corresponding to each of the V-region combinations (i.e., a total of 10,000 fragments) were sampled from the simulated microbiome. The results obtained with the combinatorial strategy were compared against the results obtained when each of the V-region combinations were targeted separately (with a sequencing depth of 10,000 reads in each case).

Results in **Table 2** indicate that although the V1+V4 and V1+V5 regions can classify the reads with commendable accuracy, the abundance values provided for individual genera deviates from the actual (RDP) lineage by a certain extent. The combinatorial approach was observed to moderate these deviations to a significant extent, and relative abundance of individual genera ascertained by the combinatorial approach exhibited better coherence with the actual lineage. In quantitative terms, while the average deviations (from actual lineage) in relative taxonomic abundance predictions for V1+V4 and V1+V5 combination–based approaches were 17.4% and 11.5%, respectively, the combinatorial approach exhibited a significantly lower average deviation (6.9%) from the actual lineage. Similar improvements were also observed when this approach was tested on microbiomes pertaining to other host-associated/ environmental niches (**Supplementary Table S9**). Given that the proposed combinatorial approach does not incur any significant additional sequencing cost and is a simple *in silico* extrapolation of the results obtained with standard pair-end sequencing, adoption of the same would be easy and would enable researchers to explore the taxonomic diversity of different environments with greater accuracy. While certain additional experimental costs for primers, multiplexing barcodes, additional PCR, and handling etc. are expected to be incurred to implement the proposed combinatorial strategy, the actual sequencing (reagents) cost, constituting the bulk of the total expenditure, remains the same. The additional pre-processing and handling efforts can at most be twice compared to the sample handling efforts needed for a single paired-end sequencing experiment. However, the potential benefits in terms of an improved taxonomic resolution are expected to outweigh any inhibitions arising due to the additional, but trivial, pre-processing and handling efforts.

CONCLUSION

The suggested protocol of targeting non-contiguously placed 16S rRNA V-regions in microbiome studies can yield better taxonomic classification accuracies without any significant additional cost/effort. A simple *in silico* combinatorial strategy further allows building consensus taxonomic profiles from multiple pair-wise combinations of V-regions, while improving accuracy in taxonomic classification. The results of the current study can serve as a guideline for future 16S rRNA amplicon–based microbiome studies and help researchers to choose the most optimal combination of V-regions for their experiment/ sampled environment.

AUTHOR CONTRIBUTIONS

AD and MH conceived the idea. NP performed the computational analysis with assistance from AD. NP, AD, MH, and SM interpreted the results and drafted the final manuscript.

FUNDING

Authors of this study are employees of Tata Consultancy Services Ltd., Pune, India. The company provided support in the form of salaries for authors, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge Mr. Deepak Yadav, TCS-Research, for his help in coding for the scripts used to create the simulated microbiome datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00653/full#supplementary-material

REFERENCES

Alekseyenko, A. V., Perez-Perez, G. I., De Souza, A., Strober, B., Gao, Z., Bihan, M., et al. (2013). Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome* 1, 31. doi: 10.1186/2049-2618-1-31

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2, e00191–16. doi: 10.1128/mSystems.00191-16

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G., and Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl. Environ. Microbiol.* 77, 3846–3852. doi: 10.1128/AEM.02772-10

Botero, L. E., Delgado-Serrano, L., Cepeda, M. L., Bustos, J. R., Anzola, J. M., Del Portillo, P., et al. (2014). Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. *Microbiome* 2, 29. doi: 10.1186/2049-2618-2-29

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

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17, 840–862. doi: 10.1128/CMR.17.4.840-862.2004

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Cui, Z., Zhou, Y., Li, H., Zhang, Y., Zhang, S., Tang, S., et al. (2012). Complex sputum microbial composition in patients with pulmonary tuberculosis. *BMC Microbiol.* 12, 276. doi: 10.1186/1471-2180-12-276

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., et al. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* 17, 55. doi: 10.1186/s12864-015-2194-9

Dutta, A., Tandon, D., Mh, M., Bose, T., and Mande, S. S. (2014). Binpairs: utilization of illumina paired-end information for improving efficiency of taxonomic binning of metagenomic sequences. *PLOS ONE* 9, e114814. doi: 10.1371/journal.pone.0114814

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., et al. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* 2, 6. doi: 10.1186/2049-2618-2-6

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

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6, 17. doi: 10.1186/s40168-017-0396-x

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6, 1176–1185. doi: 10.1038/ismej.2011.191

Hartmann, M., Howes, C. G., Abarenkov, K., Mohn, W. W., and Nilsson, R. H. (2010). V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *J. Microbiol. Methods* 83, 250–253. doi: 10.1016/j.mimet.2010.08.008

Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kato, T., Fukuda, S., Fujiwara, A., Suda, W., Hattori, M., Kikuchi, J., et al. (2014). Multiple omics uncovers host–gut microbial mutualism during prebiotic fructooligosaccharide supplementation. *DNA Res.* 21, 469–480. doi: 10.1093/dnares/dsu013

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *BioMed Res. Int.* 2012, 251364. doi: 10.1155/2012/251364

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439. doi: 10.1038/nbt.2198

Martínez-Porchas, M., Villalpando-Canchola, E., and Vargas-Albores, F. (2016). Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon* 2, e00170. doi: 10.1016/j.heliyon.2016.e00170

Moustafa, A., Li, W., Anderson, E. L., Wong, E. H. M., Dulai, P. S., Sandborn, W. J., et al. (2018). Genetic risk, dysbiosis, and treatment stratification using host genome and gut microbiome in inflammatory bowel disease. *Clin. Transl. Gastroenterol.* 9, e132. doi: 10.1038/ctg.2017.58

Munson, M. A., Banerjee, A., Watson, T. F., and Wade, W. G. (2004). Molecular analysis of the microflora associated with dental caries. *J. Clin. Microbiol.* 42, 3023–3029. doi: 10.1128/JCM.42.7.3023-3029.2004

Muscarella, M. E., Boot, C. M., Broeckling, C. D., and Lennon, J. T. (2019). Resource heterogeneity structures aquatic bacterial communities. *ISME J.* 1. doi: 10.1038/s41396-019-0427-7

Nagpal, S., Haque, M. M., and Mande, S. S. (2016). Vikodak - a modular framework for inferring functional potential of microbial communities from 16S Metagenomic Datasets. *PLOS ONE* 11, e0148347. doi: 10.1371/journal.pone.0148347

Panek, M., Paljetak, H. Č., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., et al. (2018). Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. *Sci. Rep.* 8, 5143. doi: 10.1038/s41598-018-23296-4

Petti, C. A., Polage, C. R., and Schreckenberger, P. (2005). The Role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. *J. Clin. Microbiol.* 43, 6123–6125. doi: 10.1128/JCM.43.12.6123-6125.2005

Rodrigue, S., Materna, A. C., Timberlake, S. C., Blackburn, M. C., Malmstrom, R. R., Alm, E. J., et al. (2010). Unlocking Short Read Sequencing for Metagenomics. *PLOS ONE* 5, e11840. doi: 10.1371/journal.pone.0011840

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Bieda, J., et al. (2014). The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. *Microbiome* 2, 18. doi: 10.1186/2049-2618-2-18

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., et al. (2014). Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* 80, 7583–7591. doi: 10.1128/AEM.02206-14

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

Schmalenberger, A., Schwieger, F., and Tebbe, C. C. (2001). Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. *Appl Environ Microbiol.* 67, 3557–3563. doi: 10.1128/AEM.67.8.3557-3563.2001

Soergel, D. A. W., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440–1444. doi: 10.1038/ismej.2011.208

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Xiao, S., Fei, N., Pang, X., Shen, J., Wang, L., Zhang, B., et al. (2014). A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome. *FEMS Microbiol. Ecol.* 87, 357–367. doi: 10.1111/1574-6941.12228

Yadav, D., Dutta, A., and Mande, S. S. (2019). OTUX: V-region specific OTU database for improved 16S rRNA OTU picking and efficient cross-study taxonomic comparison of microbiomes. *DNA Res.* 26, 147–156. doi: 10.1093/dnares/dsy045

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., et al. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Sci. Total Environ.* 618, 1254–1267. doi: 10.1016/j.scitotenv.2017.09.228

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

Conflict of Interests Statement: The authors are employees of the Research and Development Division of Tata Consultancy Services Ltd., Pune, India, which is a commercial company. However, this does not alter their adherence to journal policies on sharing data and materials. The authors also declare that no competing interests exist.

PERSPECTIVE
published: 09 August 2019
doi: 10.3389/fgene.2019.00721

On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications

Bartholomeus van den Bogert [1], ***Jos Boekhorst*** [2], ***Walter Pirovano*** [3] ***and Ali May*** [1]*

[1] Research and Development Dept., BaseClear, Leiden, Netherlands, [2] NIZO Food Research, Ede, Netherlands,
[3] Bioinformatics Dept., BaseClear, Leiden, Netherlands

Advances in sequencing and computational biology have drastically increased our capability to explore the taxonomic and functional compositions of microbial communities that play crucial roles in industrial processes. Correspondingly, commercial interest has risen for applications where microbial communities make important contributions. These include food production, probiotics, cosmetics, and enzyme discovery. Other commercial applications include software that takes the user's gut microbiome data as one of its inputs and outputs evidence-based, automated, and personalized diet recommendations for balanced blood sugar levels. These applications pose several bioinformatic and data science challenges that range from requiring strain-level resolution in community profiles to the integration of large datasets for predictive machine learning purposes. In this perspective, we provide our insights on such challenges by touching upon several industrial areas, and briefly discuss advances and future directions of bioinformatics and data science in microbiome research.

Keywords: DNA sequencing, microbiome, industrial biotechnology, probiotics, 16S rRNA gene profiling, metagenomics, bioinformatics, data science

OPEN ACCESS

Edited by:
Jens Stoye,
Bielefeld University, Germany

Reviewed by:
Gavin Douglas,
Dalhousie University, Canada
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

****Correspondence:***
Ali May
ali.may@baseclear.nl

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

Received: *15 April 2019*
Accepted: *09 July 2019*
Published: *09 August 2019*

Citation:
van den Bogert B, Boekhorst J, Pirovano W and May A (2019) On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications.
Front. Genet. 10:721.
doi: 10.3389/fgene.2019.00721

INTRODUCTION

Microbial communities play important roles in industrial processes such as the production of food, beverages, probiotics, paper, and cleaning products (for a review, see Singh et al., 2016). It has become an industrial standard to study the taxonomic composition and functional capabilities of these microorganisms using marker gene (e.g. 16S rRNA) and shotgun metagenome sequencing for product development, optimization, and quality control (Costessi et al., 2018). In addition, data from other omics sources such as metatranscriptomics and metabolomics can be used in integrative studies to generate leads, for instance in enzyme discovery. Some of the questions asked in these microbiome studies are related to determining the efficacy of probiotics and require strain-level characterization of the community composition (McFarland et al., 2018). Other studies focus on assessing the capability of microbial communities to produce certain compounds and necessitate recovering bacterial genomes from complex (e.g. soil) microbiomes (Howe et al., 2014). Extending microbiome applications to the public for actionable results, for example, to control blood sugar levels, requires a combination of advanced computational methods from bioinformatics, data mining, and machine learning (Zeevi et al., 2015).

In this perspective, we give an overview of several industrial microbiome applications with their bioinformatic and data science challenges. In addition, we highlight some of the advances that have the potential to provide valuable insights into the challenges facing these applications.

We conclude with sharing our view on the future directions and requirements of industrial microbiome applications in terms of their computational components.

CURRENT APPLICATIONS AND PRODUCTS

Dairy Starter Cultures

Microbial populations (e.g. of lactic acid bacteria) are used in a variety of food and beverage production processes including the manufacture of cheese, yoghurt, meat, and wine. Specifically, their role in taste and structure formation is essential, for instance during cheese ripening. These processes are governed by the presence or absence of strain-specific enzymes (Escobar-Zepeda et al., 2016). Studying such enzymes through strain isolation is often costly and time-consuming since culturing strain representatives is difficult due to laborious or unknown growth conditions (Lagier et al., 2016). Alternatively, these enzymes can be studied by metagenome sequencing, assembly, and annotation, for instance, in product optimization (De Filippis et al., 2017). In addition, metagenome assembly plays an important role in analyzing bacteriophage populations in cultures in terms of their abundance, diversity, and development (Muhammed et al., 2017), which is important not only in the prevention of phage infections that cause fermentation failures, but also for unlocking the potential of phages against food-borne pathogens (Fernández et al., 2017).

Probiotics

Probiotics are microbes that are intended to benefit the host health when consumed in adequate amounts. Identification of novel probiotics is a laborious process that starts with constructing a strain library using a culturomics approach (Lagier et al., 2016). This is followed by *in vitro* and computational research on the obtained strains to functionally characterize them, for instance for their bile resistance and potential to survive the passage of the stomach. Each of these steps reduces the list of high-potential candidates that as a final step must pass regulatory offices such as the European Food Safety Authority (EFSA, FEEDAP et al., 2018). We believe that the findings from comparative studies of the gut microbiome that highlight associations between phenotypic traits such as inflammation (Andoh et al., 2012) and obesity (Kasai et al., 2015) and specific bacterial populations, when integrated with other sources like metabolomic, demographic, dietary, and lifestyle datasets, may allow automated (e.g. machine learning-based) identification of candidate probiotic strains and reduce the time and financial cost of probiotics screening.

Small differences in the gene content of otherwise genetically identical bacterial strains can lead to different phenotypes (Zeevi et al., 2019), which in return may result in different outcomes *in vivo*. Therefore, well-conducted clinical trials are necessary to prove that the probiotic candidate itself confers the health effect. To make sure that the observed effects are not elicited by other (closely related) organisms and can be ascribed solely to the consumed probiotic, metagenomic, and bioinformatic methods that enable strain-level identification and tracking of the studied probiotic strain are required. For instance, in the genus *Bifidobacterium*, genetic differences between different strains of the same species underlie differences in carbohydrate utilization profiles (Arboleya et al., 2018). As these phenotypic traits are important in the development of probiotics for infant nutrition, applying shotgun metagenomics instead of amplicon sequencing for strain-level characterization may have substantial advantages.

Quality Control

Products like probiotics and dairy starter cultures contain live organisms that are either sold directly to consumers or used to manufacture consumer products. Next to the checks performed for raw materials, quality control of the end product is necessary to ensure the presence of correct strains and the absence of pathogens (Fenster et al., 2019). As mentioned above, microbial strains of the same species can have vastly different phenotypes, making strain-level identification in the quality control process crucial for recognizing possible contaminants (Huys et al., 2013). Traditional typing approaches such Random Amplification of Polymorphic DNA-PCR (RAPD-PCR) can be used for identifying single-strain probiotics contaminants, but require cultivation (Mohkam et al., 2016), making them unsuitable for high-throughput screening of products with complex communities (e.g. probiotics and dairy products). Whole-metagenome sequencing and analysis has the potential not only to circumvent these lengthy processes in providing strain-level information, but also to enable screening of undesired traits such as (spore) heat-resistance based on the presence of associated genes (Berendsen et al., 2016).

Cosmetics

The cosmetics industry has a growing interest in studies that aim to explore the skin microbiome as a potential therapeutic target for disorders including acne, eczema, and *Malassezia* folliculitis (Wallen-Russell, 2019). Unfortunately, these studies are typically hampered by the low biomass of skin samples, where small contaminations (e.g. from adjacent skin or reagents) can easily lead to incorrect outcomes (Kong et al., 2017). Furthermore, the human skin microbiome is strongly subject-specific (Zeeuwen et al., 2012), making it difficult to determine the effect of skin products on the general population. While this opens a potential market for personalized skin products, it also raises the need for personal longitudinal studies, where statistical methods such as redundancy analysis and principle response curve (Van den Brink and Braak, 1999) help assess correlations between taxonomic or functional composition and sample characteristics (environmental variables). Furthermore, the data can be corrected for one of the variables, such as ‘subject’ so that the variance from that covariate is removed before the actual analysis is performed, which facilitates determining the effect of the treatment.

Enzyme Discovery

A wide range of industrial enzymes, such as those used in the production of cleaning agents, laundry detergents, paper, and textile, have the continuous demand to become cheaper, greener, and more efficient. Among others, marine, soil, and lake microbiomes, with their extremely high and mainly uncharacterized biodiversity, constitute exciting functional mines not only in the search for new enzymes with such desired properties, but also for the discovery of novel enzymes that can catalyze challenging reactions (Popovic et al., 2015). A notable example of the latter is the recent discovery of two enzymes that enable the production of a renewable alternative to toluene, a petrochemical with a market of 29 million tons per year, from complex microbial communities that live in sewages and lakes (Beller et al., 2018).

Two main bioinformatic challenges in metagenomic enzyme discovery arise from the same fact that makes the chosen environment (e.g. soil) attractive in the first place: its high and uncharacterized biodiversity. The large number of different genomes in the environment and their highly skewed abundance distribution make it difficult to obtain contiguous and complete assemblies (Ayling et al., 2019), an outcome that negatively impacts gene prediction. The next challenge lies in functionally annotating the predicted genes, where commonly a high percentage of sequences are labeled as "hypothetical" or with unknown function.

Microbiome-Based Health and Personalized Nutrition

Companies and citizen science projects such as MyMicroZoo[1], Biovis[2], and American Gut[3] offer affordable microbiome analysis services to general consumers. While operationally their analyses are the same as those used for research, they must pay far more attention to the clarity of the results to ensure correct interpretations by the end-users even if the results are stated not to be interpreted as diagnosis. In practice, this means that the end-user should be guided through the (actionable) results with the help of trained healthcare professionals [e.g. dieticians and general practitioners (GPs)], who should take the limitations of a given analysis into account to prevent overinterpretation.

While basing health-related advice on published research findings is a good practice, the fact that most studies focus on a defined cohort and report "averaged" population trends makes it questionable whether results can be translated back to individuals. Such translations to the individual may be less complicated with function-based approaches through metagenomics as the 'personalized' effects are less pronounced in these datasets (Lloyd-Price et al., 2017). Nonetheless, the predictive value of a person's gut microbiome for health was demonstrated by an inspirational study by Zeevi and colleagues (2015), which integrated blood parameters, dietary habits, anthropometrics, physical activity, and the gut microbiome data into a machine learning algorithm that predicted the post meal glycemic responses of the subjects. Ultimately, 72 taxonomic or functional features of the microbiome were included in the predictive model. This approach, validated further with another independent cohort, is now offered to the public by DayTwo[4], which is a good example of how extensive datasets from scientific studies and data science can be combined in an industrial setting for providing customers with evidence-based health-related recommendations.

CURRENT ADVANCES

Metagenome Assembly, Binning, and Annotation

Metagenome assembly enables gene prediction, annotation, and abundance profiling, and therefore is an important computational step when studying the functional composition and capacity of microbiomes. Many (de Bruijn graph-based) metagenome assembly methods that differ in terms of their ease of use, scalability, running time, and memory requirement exist, making it important to carefully choose the one that serves the research question at hand the best (Van der Walt et al., 2017). For instance, in comparative studies with large cohorts where the impact of probiotics on the abundances of gene groups and pathways is analyzed, tools that are computationally less intensive, such as MEGAHIT (Li et al., 2015), are preferred. In contrast, studies with a low number of samples, such as those in enzyme discovery applications, can make use of assembly tools like metaSPAdes (Nurk et al., 2017) that include optimizations such as error correction but with a subsequent runtime trade-off. When higher read depth for assembling low abundance members or recovering full genomes is needed, data from (not too) different samples (e.g. dairy starter cultures) can be combined using co-assembly methods like crass (Dutilh et al., 2012) which also facilitates metagenomic comparison between samples. Finally, binning methods such as MetaBAT (Kang et al., 2015), MaxBin (Wu et al., 2014), and COCACOLA (Lu et al., 2017) facilitate extracting individual (draft) genomes from metagenome assemblies, which helps look at a specific organism in more detail e.g. in enzyme discovery applications where identifying the genome that encodes the target enzyme is important.

In a recent study of cow rumen microbiome, a valuable environment for biomass-degrading enzyme discovery, Stewart et al. (2018) showed that 90% of the proteins predicted to be involved in the studied mechanism (carbohydrate metabolism) did not have a good match in public databases. Such findings highlight the relatively large room for improvement in microbiome annotation.

Hypothesis-Driven Functional Analyses

Exhaustively analyzing all functional aspects and querying all potential longitudinal and cross-sectional aspects of a microbiome dataset is generally considered a hopeless task. Even when computationally feasible, multiple testing issues lead to a severe reduction of the analysis power. Although approaches like the removal of collinear variables and validation of potential correlations in independent datasets can in part address these issues (Falony et al., 2016), delineating the relevant functional aspects is a big step in overcoming these limitations. Using a specific database to answer a particular hypothesis, such as in the case with certain enzyme classes or a set of enzymatic pathways, is such an approach. Examples of such databases

[1] mymicrozoo.com
[2] biovis-diagnostik.eu
[3] humanfoodproject.com/americangut/
[4] daytwo.com

and tools are Resfams (Gibson et al., 2015), dbCAN (Yin et al., 2012), and antiSMASH (Blin et al., 2017), focusing on antibiotic resistance, carbohydrate utilization, and secondary metabolite synthesis, respectively. Methods developed for the elucidation of gene function, such as the guilt by association approaches implemented in STRING (Szklarczyk et al., 2014), can be used to identify genes that are not directly flagged by comparison to specific functional datasets such as the ones described above, but have distribution patterns similar enough to genes that are represented in the reference set. A drawback of functional analyses that require protein sequences is the need for assembly and gene prediction, which can be computationally intensive as described above. Tools like HUMAnN2 (Franzosa et al., 2018) work directly with short-read data without requiring an assembly for profiling protein family abundance.

Assembly-Independent Strain-Level Characterization

Probiotic members such as *Bifidobacterium longum* subsp. *longum* and *Bifidobacterium longum* subsp. *infantis*, which have two distinct phenotypes with relevant functional implications in infant nutrition (Underwood et al., 2015), differ only slightly in their16S rRNA gene sequences (Lawley et al., 2017). Such differences are lost in classical operational taxonomic unit (OTU) clustering-based taxonomic analyses. Novel methods like UNOISE2 (Edgar, 2016) and DADA2 (Callahan et al., 2016) circumvent clustering and apply sequence filtering steps, enabling distinguishing between sequences on a single-nucleotide level by grouping reads in amplicon sequence variants (ASVs). This has a great potential to improve the phylogenetic depth at which microbiome studies can be interpreted. Notable applications of these new algorithms provided new, sub-species level insights into oral (Mukherjee et al., 2018) and vaginal microbiomes (Callahan et al., 2017).

In cases where multiple strains of a species of interest have identical 16S rRNA sequences, algorithms such as StrainPhlAn (Truong et al., 2017) and PanPhlAn (Scholz et al., 2016) enable strain-level analyses from shotgun metagenome datasets without the need for metagenome assembly (**Figure 1**). These methods open the possibility for routine compositional analyses to verify the presence of desired strains or identify potential pathogens in end products.

Long-Read Sequencing and Other Advances

Although their use in microbiome studies is currently not common, long-read sequencing platforms Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) offer exciting opportunities for several industrial applications mentioned above. For instance, circular consensus sequencing application by PacBio, which allows multiple reads generated from a circularized amplicon molecule to be bioinformatically combined into a high-quality, full-length (16S) sequence (Callahan et al., 2018), provides the necessary phylogenetic resolution for applications such as fermentation studies, which is unfeasible with short-read amplicon sequencing. The on-demand sequencing nature of ONT, on the other hand, seems suitable for quality control applications for detecting distinct pathogens, although the high error rate is limiting for accurate, strain-level detection.

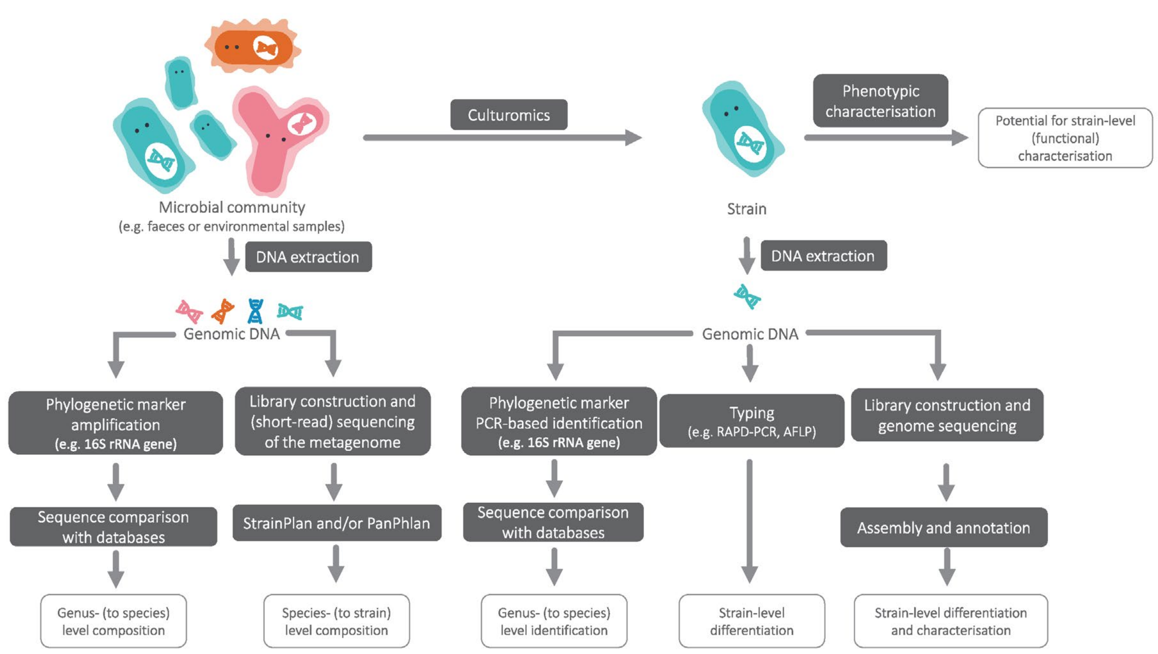

Culturomics
Phenotypic characterisation
Potential for strain-level (functional) characterisation
Microbial community
(e.g. faeces or environmental samples)
Strain
DNA extraction
DNA extraction
Genomic DNA
Genomic DNA
Phylogenetic marker amplification
(e.g. 16S rRNA gene)
Library construction and (short-read) sequencing of the metagenome
Phylogenetic marker PCR-based identification
(e.g. 16S rRNA gene)
Typing
(e.g. RAPD-PCR, AFLP)
Library construction and genome sequencing
Sequence comparison with databases
StrainPlan and/or PanPhlan
Sequence comparison with databases
Assembly and annotation
Genus- (to species) level composition
Species- (to strain) level composition
Genus- (to species) level identification
Strain-level differentiation
Strain-level differentiation and characterisation

FIGURE 1 | An overview of approaches to achieve taxonomic resolution at different levels.

Even with high dataset coverage and advanced methods, assemblies from short-read datasets commonly remain very fragmented, especially in samples from complex communities like soil. Soon, we expect the integration of long-read sequencing to be more common in assembly-oriented studies for obtaining full, chromosome-level microbial genomes. Correspondingly, we see potential in adapting hybrid assembly methods such as hybridSPAdes (Antipov et al., 2015) to enable their use with long- and short-read metagenome datasets. Other promising developments revolve around using barcoded short reads that have long-range information, such as those provided by 10x Genomics (http://10xgenomics.com), in microbiome research. We see the emergence of tailored bioinformatic methods such as the Athena assembler (Bishara et al., 2018), which uses barcode information in short-reads and improves the contiguity of metagenome assemblies.

Machine Learning and Data Science

With decreasing sequencing costs, the size of datasets in microbiome studies and the depth of sequencing per sample have increased. This led to studies with higher statistical power, and consequently to the transition of OTU tables and functional profiles from end-goal deliverables into starting material for downstream analyses such as machine learning (ML) applications (Pasolli et al., 2016). Methods like random forests (RF) have been successfully used by many within a disease context, for instance, for accurately predicting irritable bowel syndrome (Saulnier et al., 2011) and bacterial vaginosis (Beck and Foster, 2014) based on taxonomic profiles (for a review, see LaPierre et al., 2019 and Qu et al., 2019). On the other hand, Sze and Schloss (2016) used 10 previously published obesity datasets and showed that RF ML models trained on one of the datasets and tested on the remaining nine had a median accuracy of only 56.68%, suggesting that i) the method may not be applicable for some diseases, or ii) the disease signal may be more apparent at the level of differentially expressed functions (gene transcripts) of the microbiome.

Industrial microbiome applications of ML include building classification models based on soil microbiome data for detecting oil sites (Miranda et al., 2019) and above-mentioned personalized health-related lifestyle (diet) recommendation services that are partly based on gut microbiome data. As mentioned in *Probiotics*, we expect dataset integration and ML to have an impact also on areas such as screening of novel probiotics. To meet the overall demand for user-friendly ML in microbiome research, software suites like QIIME 2 (Bokulich et al., 2018), MicrobiomeAnalyst (Dhariwal et al., 2017), and USEARCH (Edgar, 2010) started incorporating ML methods that can be used by researchers who aren't necessarily trained as bioinformaticians.

CONCLUSIONS AND OUTLOOK

The vast number of experimental and computational methods available for microbiome research have led to a broad collection of choices. While creation of guidelines and standardization for increased comparability and reproducibility is essential, achieving a global consensus in methods used remains a challenge. What constrains researchers to their current practices is mainly the laborious nature of adopting other (new) protocols, which may have an ironically detrimental effect on comparability between different studies, or even within studies that run over prolonged periods. Like Knight et al. (2018), we think that a primary objective of microbiome studies should be to standardize the documentation of used methods, tools, data formats, and data processing parameters, and to publish these "logs" next to the final results and interpretations. While complete disclosure is scientifically ideal, it raises commercial concerns for microbiome analysis providers like BaseClear[5], NIZO food research[6], Clinical Microbiomics[7], Vedanta Biosciences[8], and COSMOSID[9], as it would mean releasing a substantial part of their, sometimes unique, intellectual property.

With reducing costs, we soon expect long-read sequencing technologies to be commonly used in microbiome studies, which will benefit from enhanced taxonomic resolution with full-length marker gene sequencing, as well as improved functional analyses thanks to more contiguous metagenome assemblies. Here, the focus in developments is likely to be on the translation of bioinformatic protocols already established for short reads to long-read versions, for instance in denoising and read classification approaches.

Other challenges relate to shotgun metagenome analyses in large studies, where expensive calculations used in *de novo* assembly and annotation may result in capacity issues. For companies that cannot afford large on-premise compute infrastructures, the cloud provides a flexible alternative, where know-how of cloud-computing becomes essential.

Finally, the rapid translation of microbiome research into important industrial applications in healthcare, energy, and food production will continue to stimulate collaborations between academic and industrial communities. We expect the role of bioinformatics and data science to become only more significant in this relationship.

AUTHOR CONTRIBUTIONS

All authors were involved in the writing and final preparation of the article.

ACKNOWLEDGMENTS

The authors thank Thomas Battaglia for his help in the preparation of the final manuscript.

[5] baseclear.com
[6] nizo.com
[7] clinical-microbiomics.com
[8] vedantabio.com
[9] cosmosid.com

REFERENCES

(FEEDAP), E., Rychen, G., Aquilina, G., Azimonti, G., Bampidis, V., Bastos, M., et al. (2018). Guidance on the characterisation of microorganisms used as feed additives or as production organisms. *EFSA J.* 16 (3), e05206. doi: 10.2903/j.efsa.2018.5206

Andoh, H., Kizuoka, H., Tsujikawa, T, Nakamura, S, Hirai, F., Suzuki, Y., et al. (2012). Multicenter analysis of fecal microbiota profiles in Japanese patients with Crohn's disease. *J. Gastroenterol.* 47 (12), 1298–1307. doi: 10.1007/s00535-012-0605-0

Antipov, D., Korobeynikov, A., McLean, J., and Pevzner, P. (2015). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32 (7), 1009–1015. doi: 10.1093/bioinformatics/btv688

Arboleya, S., Bottacini, F., O'Connell-Motherway, M., Ryan, C., Ross, R., Van Sinderen, D., et al. (2018). Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* 19 (1), 33. doi: 10.1186/s12864-017-4388-9

Ayling, M., Clark, M., and Leggett, R. (2019). New approaches for metagenome assembly with short reads. *Brief Bioinform.* 1–11. doi: 10.1093/bib/bbz020

Beck, D., and Foster, J. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PloS One* 9 (2), e87830. doi: 10.1371/journal.pone.0087830

Beller, H., Rodrigues, A., Zargar, K., Wu, Y.-W., Saini, A., Saville, R., et al. (2018). Discovery of enzymes for toluene synthesis from anoxic microbial communities. *Nat. Chem. Biol.* 14 (5), 451. doi: 10.1038/s41589-018-0017-4

Berendsen, E., Boekhorst, J., Kuipers, O., and Wells-Bennik, M. (2016). A mobile genetic element profoundly increases heat resistance of bacterial spores. *ISME J.* 10 (11), 2633. doi: 10.1038/ismej.2016.59

Bishara, A., Moss, E., Kolmogorov, M., Parada, A., Weng, Z., Sidow, A., et al. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* 36, 1067–1075. doi: 10.1038/nbt.4266

Blin, K., Wolf, T., Chevrette, M., Lu, X., Schwalen, C., Kautsar, S., et al. (2017). antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* 45 (W1), W36–W41. doi: 10.1093/nar/gkx319

Bokulich, N., Dillon, M., Bolyen, E., Kaehler, B., Huttley, G., and Caporaso, J. (2018). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.* 3 (30), 934. doi: 10.21105/joss.00934

Callahan, B., DiGiulio, D., Goltsman, D., Sun, C., Costello, E., Jeganathan, P., et al. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc. Natl. Acad. Sci* 114 (37), 9966–9971. doi: 10.1073/pnas.1705899114

Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., and Holmes, S. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13 (7), 581. doi: 10.1038/nmeth.3869

Callahan, B., Wong, J., Heiner, C., Oh, S., Theriot, C., Gulati, A., et al. (2018). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* gkz569. https://doi.org/10.1093/nar/gkz569

Costessi, A., van den Bogert, B., May, A., Ver Loren van Themaat, E., Roubos, J., Kolkman, M., et al. (2018). Novel sequencing technologies to support industrial biotechnology. *FEMS Microbiol. Letters* 365 (16), fny103. doi: 10.1093/femsle/fny103

De Filippis, F., Parente, E., and Ercolini, D. (2017). Metagenomics insights into food fermentations. *Microb. Biotechnol.* 10 (1), 91–102. doi: 10.1111/1751-7915.12421

Dhariwal, A., Chong, J., Habib, S., King, I., Agellon, L., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45 (W1), W180–W188. doi: 10.1093/nar/gkx295

Dutilh, B., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R., et al. (2012). Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28 (24), 3225–3231. doi: 10.1093/bioinformatics/bts613

Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26 (19), 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv* 081257. doi: 10.1101/081257

Escobar-Zepeda, A., Sanchez-Flores, A., and Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127. doi: 10.1016/j.fm.2016.02.004

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352 (6285), 560–564. doi: 10.1126/science.aad3503

Fenster, K., Freeburg, B., Hollard, C., Wong, C., Rønhave Laursen, R., and Ouwehand, A. (2019). The production and delivery of probiotics: a review of a practical Approach. *Microorganisms* 7 (3), 83. doi: 10.3390/microorganisms7030083

Fernández, L., Escobedo, S., Gutiérrez, D., Portilla, S., Martínez, B., García, P., et al. (2017). Bacteriophages in the dairy environment: from enemies to allies. *Antibiotics* 6 (4), 27. doi: 10.3390/antibiotics6040027

Franzosa, E., McIver, L., Rahnavard, G., Thompson, L., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15 (11), 962. doi: 10.1038/s41592-018-0176-y

Gibson, M., Forsberg, K., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 9 (1), 207. doi: 10.1038/ismej.2014.106

Howe, A., Jansson, J., Malfatti, S., Tringe, S., Tiedje, J., and Brown, C. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci.* 111 (13), 4904–4909. doi: 10.1073/pnas.1402564111

Huys, G., Botteldoorn, N., Delvigne, F., De Vuyst, L., Heyndrickx, M., Pot, B., et al. (2013). Microbial characterization of probiotics–Advisory report of the Working Group "8651 Probiotics" of the Belgian Superior Health Council (SHC). *Mol. Nutr. Food Res.* 57 (8), 1479–1504. doi: 10.1002/mnfr.201300065

Kang, D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165. doi: 10.7717/peerj.1165

Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., et al. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterol.* 15 (1), 100. doi: 10.1186/s12876-015-0330-2

Knight, R., Vrbanac, A., Taylor, B., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 4 (16), 410–422. doi: 10.1038/s41579-018-0029-9

Kong, H., Andersson, B., Clavel, T., Common, J., Jackson, S., Olson, N., et al. (2017). Performing skin microbiome research: a method to the madness. *J. Investig. Dermatol.* 137 (3), 561–568. doi: 10.1016/j.jid.2016.10.033

Lagier, J.-C., Khelaifia, S., Alou, M., Ndongo, S., Dione, N., Hugon, P., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* 1 (12), 16203. doi: 10.1038/nmicrobiol.2016.203

LaPierre, N., Ju, C.-T., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods.* doi: 10.1016/j.ymeth.2019.03.003

Lawley, B., Munro, K., Hughes, A., Hodgkinson, A., Prosser, C., Lowry, D., et al. (2017). Differentiation of Bifidobacterium longum subspecies longum and infantis by quantitative PCR using functional gene targets. *PeerJ.* 5, e3375. doi: 10.7717/peerj.3375

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 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550 (7674), 61. doi: 10.1038/nature23889

Lu, Y., Chen, T., Fuhrman, J., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* 33 (6), 791–798. doi: 10.1093/bioinformatics/btw290

McFarland, L., Evans, C., and Goldstein, E. (2018). Strain-specificity and disease-specificity of probiotic efficacy: a systematic review and meta-analysis. *Front. Med.* 5, 124. doi: 10.3389/fmed.2018.00124

Miranda, J., Seoane, J., Esteban, A., and Espí, E. (2019). *Microbial Exploration Techniques: An Offshore Case Study.* Eds. J. Miranda, J. Seoane, A. and Esteban, E. Espí. Boca Raton, Florida: CRC Press

Mohkam, M., Nezafat, N., Berenjian, A., Mobasher, M., and Ghasemi, Y. (2016). Identification of Bacillus probiotics isolated from soil rhizosphere using 16S rRNA, recA, rpoB gene sequencing and RAPD-PCR. *Probiotics Antimicrob. Proteins* 8 (1), 8-18. doi: 10.1007/s12602-016-9208-z

Muhammed, M., Kot, W., Neve, H., Mahony, J., Castro-Mejía, J., Krych, L., et al. (2017). Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. *Appl. Environ. Microbiol.* 83 (19), e00888–e00817. doi: 10.1128/AEM.00888-17

Mukherjee, C., Beall, C., Griffen, A., and Leys, E. (2018). High-resolution ISR amplicon sequencing reveals personalized oral microbiome. *Microbiome* 6 (1), 153. doi: 10.1186/s40168-018-0535-z

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27 (5), 824–834. doi: 10.1101/gr.213959.116

Pasolli, E., Truong, D., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12 (7), e1004977. doi: 10.1371/journal.pcbi.1004977

Popovic, A., Tchigvintsev, A., Tran, H., Chernikova, T., Golyshina, O., Yakimov, M., et al. (2015). *Metagenomics as a tool for enzyme discovery: hydrolytic enzymes from marine-related metagenomes*. Eds. A. Popovic, A. Tchigvintsev, H. Tran, T. Chernikova, O. Golyshina, and M. Yakimov. Basel, Switzerland: Springer. doi: 10.1007/978-3-319-23603-2_1

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of Machine Learning in Microbiology. *Front. Microbiol.* 10, 827. doi: 10.3389/fmicb.2019.00827

Saulnier, D., Riehle, K., Mistretta, T.–A., Diaz, M.–A., Mandal, D., Raza, S., et al. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterol.* 141 (5), 1782–1791. doi: 10.1053/j.gastro.2011.06.072

Scholz, M., Ward, D., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13 (5), 435. doi: 10.1038/nmeth.3802

Singh, R., Kumar, M., Mittal, A., and Mehta, P. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech.* 6 (2), 174. doi: 10.1007/s13205-016-0485-8

Stewart, R., Auffret, M., Warr, A., Wiser, A., Press, M., Langford, K., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* 9 (1), 870. doi: 10.1038/s41467-018-03317-6

Sze, M., and Schloss, P. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio.* 7 (4), e01018–e01016. doi: 10.1128/mBio.01018-16

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43 (D1), D447–D452. doi: 10.1093/nar/gku1003

Truong, D., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27 (4), 626–638. doi: 10.1101/gr.216242.116

Underwood, M., German, J., Lebrilla, C., and Mills, D. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res.* 77 (1-2), 229. doi: 10.1038/pr.2014.156

Van den Brink, P., and Braak, C. (1999). Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environ. Toxicol. Chem. Int J.* 18 (2), 138–148. doi: 10.1002/etc.5620180207

Van der Walt, A., Van Goethem, M., Ramond, J.-B., Makhalanyane, T., Reva, O., and Cowan, D. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18 (1), 521. doi: 10.1186/s12864-017-3918-9

Wallen-Russell, C. (2019). The Role of Every-Day Cosmetics in Altering the Skin Microbiome: a Study Using Biodiversity. *Cosmetics* 6 (1), 2. doi: 10.3390/cosmetics6010002

Wu, Y.-W., Tang, Y.-H., Tringe, S., Simmons, B., and Singer, S. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome.* 2 (1), 26. doi: 10.1186/2049-2618-2-26

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 40 (W1), W445–W451. doi: 10.1093/nar/gks479

Zeeuwen, P., Boekhorst, J., van den Bogaard, E., de Koning, H., van de Kerkhof, P., Saulnier, D., et al. (2012). Microbiome dynamics of human epidermis following skin barrier disruption. *Genome Biol.* 13 (11), R101. doi: 10.1186/gb-2012-13-11-r101

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., et al. (2019). Structural variation in the gut microbiome associates with host health. *Nature* 1, 43–48. doi: 10.1038/s41586-019-1065-y

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163 (5), 1079–1094. doi: 10.1016/j.cell.2015.11.001

Conflict of Interest Statement: BB, WP, and AM work at BaseClear. JB works at NIZO Food Research.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 August 2019
doi: 10.3389/fgene.2019.00753

MetaTOR: A Computational Pipeline to Recover High-Quality Metagenomic Bins From Mammalian Gut Proximity-Ligation (meta3C) Libraries

Lyam Baudry[1,2,3†], Théo Foutel-Rodier[1,2,3†], Agnès Thierry[1,2], Romain Koszul[1,2] and Martial Marbouty[1,2*]*

[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR3525, CNRS, Paris, France, [2] Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Paris, France, [3] Sorbonne Université, Collège Doctoral, Paris, France

OPEN ACCESS

Edited by:
Bas E. Dutilh,
Utrecht University, Netherlands

Reviewed by:
Marnix H. Medema,
Wageningen University & Research, Netherlands
Digvijay Verma,
Babasaheb Bhimrao Ambedkar University, India

***Correspondence:**
Romain Koszul
romain.koszul@pasteur.fr
Martial Marbouty
martial.marbouty@pasteur.fr

†These authors have contributed equally to this work.

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

Received: *17 April 2019*
Accepted: *17 July 2019*
Published: *20 August 2019*

Citation:
Baudry L, Foutel-Rodier T, Thierry A, Koszul R and Marbouty M (2019) MetaTOR: A Computational Pipeline to Recover High-Quality Metagenomic Bins From Mammalian Gut Proximity-Ligation (meta3C) Libraries. Front. Genet. 10:753. doi: 10.3389/fgene.2019.00753

Characterizing the complete genomic structure of complex microbial communities would represent a key step toward the understanding of their diversity, dynamics, and evolution. Current metagenomics approaches aiming at this goal are typically done by analyzing millions of short DNA sequences directly extracted from the environment. New experimental and computational approaches are constantly sought for to improve the analysis and interpretation of such data. We developed MetaTOR, an open-source computational solution that bins DNA contigs into individual genomes according to their 3D contact frequencies. Those contacts are quantified by chromosome conformation capture experiments (3C, Hi-C), also known as proximity-ligation approaches, applied to metagenomics samples (meta3C). MetaTOR was applied on 20 meta3C libraries of mice gut microbiota. We quantified the program ability to recover high-quality metagenome-assembled genomes (MAGs) from metagenomic assemblies generated directly from the meta3C libraries. Whereas nine high-quality MAGs are identified in the 148-Mb assembly generated using a single meta3C library, MetaTOR identifies 82 high-quality MAGs in the 763-Mb assembly generated from the merged 20 meta3C libraries, corresponding to nearly a third of the total assembly. Compared to the hybrid binning softwares MetaBAT or CONCOCT, MetaTOR recovered three times more high-quality MAGs. These results underline the potential of 3C-/Hi-C-based approaches in metagenomic projects.

Keywords: metagenomics Hi-C, gut microbiome, Hi-C, metagenomics binning, metagenomic analysis, binning algorithm, metagenome-assembled genomes

INTRODUCTION

Microbial communities hold important roles in ecosystems regulation (Philippot et al., 2013; Edbeib et al., 2016; Coutinho et al., 2018; Rosado et al., 2019), such as the human gut (Cho and Blaser, 2012). Understanding the behaviors of these communities is a complex task, and one important step toward this objective relies on the characterization of the genomes of the different species within (Long et al., 2016). Indeed, the genome sequence allows to infer metabolic pathways and, by extension, provide indications about the species lifestyle in the environment. Supported by high-throughput sequencing technologies dropping costs and backed by increasingly powerful computational resources, the field of metagenomics aims at exploring ecosystems through the analysis of DNA sequences extracted directly from the environment to gain insights on microbial population diversity and dynamics (Spang et al., 2015; Hug et al., 2016; Paez-Espino et al., 2016; Castelle and Banfield, 2018). Characterizing complete or near-complete genomes remains however difficult to achieve, depending to some extent to the popularity and complexity of the ecosystem studied (Olson et al., 2017; Quince et al., 2017; Sieber et al., 2018). An important aspect of metagenomics studies therefore consists in developing computation approaches to characterize genomes in metagenomics data (Albertsen et al., 2013; Alneberg et al., 2014; Frank et al., 2016; Sieber et al., 2018).

Most computational approaches rely on the composition and/or co-abundance of sequences recovered from multiple samples to pool (bin) them together (Alneberg et al., 2014; Wu et al., 2014; Kang et al., 2015; Lu et al., 2017; Graham et al., 2017; Laczny et al., 2017). Composition-based method groups together sequences that display similar metrics, such as GC content and/or tetra- and/or penta-nucleotide frequencies. Co-abundance-based approaches trace the relative amount of sequences over multiple samples and group together those with similar coverage variation. Co-abundance is very effective when multiple samples of the same ecosystem are available under different conditions. Today, most metagenomics binning pipeline consists in hybrid approaches combining both strategies to improve the confidence of the resulting sequences bins (Alneberg et al., 2014; Wu et al., 2014; Kang et al., 2015; Graham et al., 2017; Lu et al., 2017). However, caveats and limitations remain. First, grouping sequences based on their similarities imply a strong assumption regarding the homogeneity of the genomes' composition. This hypothesis is therefore not valid when horizontal transfer or introgression of genetic material takes place between species with (highly) divergent sequence compositions. For instance, the GC content of prophages and of their bacterial genomes host can differ widely. Co-abundance-based methods require multiple samples and large amounts of data to be fully effective, which can be impractical and/or costly. In addition, if several multiple species share the same genetic elements, co-abundance-based methods will also fail to identify the association of these elements with the different species.

Novel technologies, such as single-cell (Ji et al., 2017), long reads (Frank et al., 2016) or proximity ligation/chromosome conformation capture (3C) (reviewed in Marbouty and Koszul, 2015; Flot et al., 2015), hold the potential to address some of these limitations. The latter approach, dubbed meta3C from the original 3C approach (Dekker et al., 2002), aims at quantifying and exploiting collisions between DNA loci over a population of species to identify those that share the same cellular compartment. Sequences belonging to the same genome display enriched contact frequencies compared to those belonging to different genomes, as shown by applying meta3C on controlled mixes of species (Burton et al., 2014; Beitel et al., 2014; Marbouty et al., 2014). Besides controlled mixes, meta3C successfully reconstructed genomes from truly unknown and complex ecosystems as well (Marbouty et al., 2014; Marbouty et al., 2017; Stewart et al., 2018). Not only near-complete genomes from microorganisms can be recovered from a single experiment, but additional information about the genomic structure of these microbial populations can be recovered as well, including plasmids (Marbouty et al., 2014; Press et al., 2017; Stalder et al., 2019) and phage-host infection spectrum (Marbouty et al., 2017). These studies suggest that meta3C and similar approaches hold the potential to 1) accurately bin genomes and episomal DNA molecules and 2) assign episomal DNA molecules to their respective hosts. However, comprehensive, end-to-end computational pipelines to process raw meta3C datasets remain sparse (Marbouty et al., 2017; DeMaere and Darling, 2019). Most analyses so far have focused on single mock communities, and quantifiable metrics are lacking to see how meta3C-like approaches truly compare—and possibly complement—traditional binning methods, notably regarding the quality, completeness, and accuracy of retrieved bins.

To address this need, we developed MetaTOR (Metagenomic Tridimensional Organisation–based Reassembly), a lean and scalable tool to investigate single or multiple proximity-ligation (i.e., 3C or Hi-C libraries) metagenomic experiments, from raw 3C reads and assembly to bins. MetaTOR was applied on 20 meta3C libraries of mouse gut samples collected over time. This first dynamic meta3C study allowed us to reconstruct dozens of complete genome sequences, and to compare the genomic bins recovered using MetaTOR with bins generated by binning software MetaBAT (Kang et al., 2015) and CONCOCT (Alneberg et al., 2014). MetaTOR compared favorably with respect to the number of high-quality genomes recovered (Bowers et al., 2017) and the amount of binned sequences. In addition, 3C-based binning was less dependent on the quality of the metagenome assembly (in terms of fragmentation—i.e., contigs' mean size, N50). Overall, MetaTOR is a robust tool to process proximity-ligation sequencing data, regardless the number of samples processed.

MATERIALS AND METHODS

Feces Sampling and meta3C Library Generation

The feces of three groups of two mice were sampled over 20 days as follows: days 2, 5, and 9 for cage n°1; days 2, 4, 5, 6, 7, 9, 10, 12, and 16 for cage n°2; and days 2, 5, 6, 7, 9, 11, 12, and 16 for cage n°3 (**Supplementary Figure 1**). The samples were immediately cross-linked after sampling in 30 ml of 1X tris-EDTA buffer supplemented with 3% formaldehyde (final concentration), for 1 h at room temperature with agitation. Formaldehyde was quenched by adding 10 ml of 2.5 M glycine during 20 min at room temperature with moderate agitation. Samples were then recovered by centrifugation, and pellets were stored at −80°C until processing. The libraries were then prepared and sequenced using pair-end (PE) Illumina sequencing (2 × 75 bp NextSeq) as described (Marbouty et al., 2014; Foutel-Rodier et al., 2018).

Read Processing and Assembly

The first 10 bp of each read correspond to custom-made amplification primers allowing to remove PCR duplicates from the read pool (Marbouty et al., 2015). Those 10 bp were removed afterwards, and the resulting 65-pb sequences were filtered and trimmed using cutadapt (Martin, 2011). Quality was controlled with FastQC, and a total of 813 million PE reads were kept in total (over the 20 samples). Reads from libraries sampled from 1) cage 3 at day 2, 2) cage 3 with all samples, and 3) all cages with all samples were then used to perform three independent assemblies using MEGAHIT v1.1.1.2 (Li et al., 2015) with default parameters. Contigs under 500 bp were discarded from further analyses.

Assemblies Analysis

Contigs from the three assemblies were analyzed with the MG-RAST pipelines (Meyer et al., 2008). The metagenomics RAST server allows automated annotations of complete or draft microbial genomes and provides information on phylogenetic and functional classification of the contigs. It also provides an alpha diversity measurement of the assembly.

Alignment Step and Network Generation

Filtered reads were aligned independently in single-end mode using Bowtie2 v2.2.9 (option—very-sensitive-local) against one of the assemblies. For each sample, both alignment files were sorted and merged using the SAMtools and pysam libraries. Ambiguous alignments and alignments with mapping quality under 20 were discarded. All pairs of reads for which both reads aligned unambiguously on two different contigs were kept to generate the network. Contigs were considered as nodes, and the values of the edges (i.e., the weight) of the network were determined by counting the number of non-ambiguous alignments bridging the corresponding two contigs. Normalization was computed by dividing the edge value by the geometric mean of the nodes' coverage (i.e., contigs' coverage). Contig coverage was calculated using MetaBAT 1 v0.32.5 script: jgi_summarize_bam_contig_depths with a contig size limit of 500 bp for every set of reads.

Louvain Clustering

We showed before that the updated implementation of the Louvain community method provided in (Blondel et al., 2008) was a promising approach to identify subnetworks of contigs in the meta3C network that display enriched contacts between themselves (Marbouty et al., 2014). The Louvain algorithm was run 400 times on each network, using the classical Newman-Girvan criterion. Nodes that systematically clustered together for each of the first 100 iterations were pooled together in core communities (CCs), as described previously (Marbouty et al., 2017).

CCs Validation/Evaluation and Taxonomic Annotation

CCs above 500 kb were evaluated for completeness and contamination using CheckM version 1.0.7 (Parks et al., 2015). A CC was validated as a bin if its contamination rate range under 10%. CheckM was also used to assign taxa, at the class level, to validated bins using the *lineage* workflow.

MAGs Evaluation

Validated bins were further evaluated following the standards to classify MAGs as high quality, medium quality, or low quality (Bowers et al., 2017). tRNA were searched with tRNAscan-SE 2.0 (Lowe and Eddy, 1997) (option -B). 16S and 23S rRNAs were searched using METAXA2 (Bengtsson-Palme et al., 2015)(options: -g SSU and -g LSU, respectively). We used RNAmmer-1.2 (Lagesen et al., 2007) (options: -S bacteria -m tsu) to look for 5S RNA. Bins were considered high-quality draft if they had 18 or more different tRNAs and at least one of each rRNA gene.

Recursive Louvain Clustering

Partially complete CCs (> 70% completion) with contamination levels upper than 10% were selected for recursive binning. Briefly, the partition step was re-run 10 times on these contaminated CCs (i.e., on their corresponding sub-network), yielding groups of smaller CC (i.e., sub-CCs) which were then re-processed in the binning step to assess for their quality.

Pipeline Comparison

CONCOCT v1.0.0 (Alneberg et al., 2014) and MetaBAT 1 v0.32.5 (Kang et al., 2015) were run on the same set of reads and assemblies, using the different time samples for differential coverage. Resulting bins above 500 kb were retrieved and compared with MetaTOR's for completeness and contamination using CheckM. CONCOCT was run with the following parameters –r 65 -s 100. MetaBAT 1 was run with default parameters.

RESULTS

Algorithmic Principles Underneath the MetaTOR Pipeline

MetaTOR (https://github.com/koszullab/metaTOR) aims at providing the most accurate overview of genome content of a population, starting from as little as one meta3C library, while taking full advantage of additional libraries if available. It's structured around four main steps: alignment, partition, annotation, and binning (**Figure 1**). MetaTOR was purposely designed to maintain a high level of modularity and flexibility, so that users can supply their own intermediary inputs and tweak parameters to their liking at every step. This can save both time and resources. If starting from the raw data, all needed is the meta3C PE files and an assembly of the microbial community obtained either directly from the meta3C reads (as described in this work and in Marbouty et al., 2014; Marbouty et al., 2017) or from a DNA library generated independently (**Figure 1A**).

- **[Align]** (**Figure 1B**): First, meta3C reads are aligned independently along the contigs of the metagenome assembly using Bowtie2 (as aligners tend to leave out far-off alignments when run in PE mode). Contigs are then sorted, filtered for mapping quality, and merged into a global alignment file. The alignment is converted into a contact network stored in a plain text file [network.txt: column 1—node 1/column 2—node

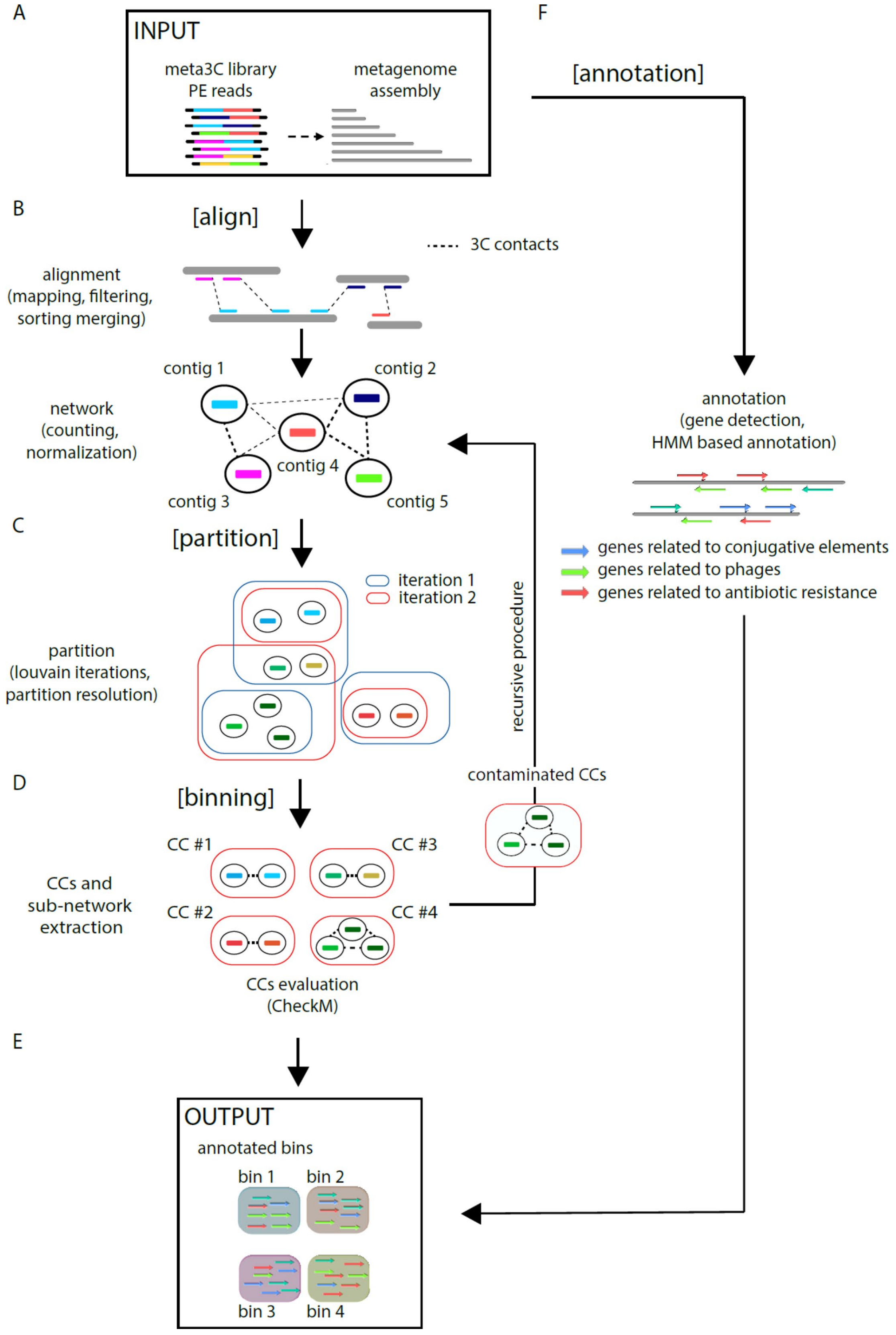


FIGURE 1 | Continued

FIGURE 1 | MetaTOR pipeline. Schematic representation of the MetaTOR pipeline. **(A)** MetaTOR is initialized with an assembly and a set of 3C/Hi-C PE reads. **(B)** [Align] will align, sort, and merge reads to deliver a network of contig interactions. **(C)** [Partition] will deconvolve the previously defined network using a Louvain iterative procedure and **(D)** [Binning] will retrieve CCs (FASTA file and corresponding sub-network) from selected partition to evaluate them using CheckM. At this step, it is possible to perform a recursive procedure on selected CCs to split them further into sub-CCs. **(F)** [Annotation] is an optional step that use HMM models to provide final annotations. **(E)** The final output of the pipeline is a set of annotated bins.

2/column 3—weight] to facilitate further third-party analysis. In the network, each node represents one contig, and each edge (a.k.a. weight) represents the contact score found between two contigs. This step integrates variable parameters such as enforcing a lower size limits for contigs or a normalization step. Normalization of the network typically uses contig coverage, but other normalizations can be implemented as well.

- [**Partition**] (**Figure 1C**): An iterative Louvain procedure is applied on the network file to partition the network into groups of contigs that consistently cluster together, i.e., "see" each other's in space more often than their neighbors' (Blondel et al., 2008; Marbouty et al., 2014; Marbouty et al., 2017). These clusters or CC constitute the matrix of the metagenomic binning. The number of iterations is a free parameter of the pipeline and can be set by the user. However, we noted that the number of CC stabilizes after a while with small oscillations around a fixed value, and therefore recommend enough cycles to reach that threshold.
- [**Binning**] (**Figure 1D**) CCs are then extracted (FASTA files) and their gene content assessed for completeness and contamination using CheckM (Parks et al., 2015). In parallel, the pipeline extracts sub-networks for each CC (i.e., network between the corresponding contigs). Extraction of each sub-network allows the user to perform, if needed, a recursive procedure at this step on the defined contig group (i.e., CCs) (see **Figure 1**—"recursive procedure"). Indeed, some CCs exhibit both a high completion rate and a high contamination levels suggesting that they contain more than one genome. By applying the partition step only on their corresponding sub-network, it becomes possible to sub-partition using the Louvain algorithm these CCs into smaller ones (i.e., sub-CCs). This step typically breaks down the most contaminated CCs into smaller, low-contaminated sub-CCs. The retrieved sub-CCs can also be evaluated using CheckM and validated as bins.
- [**Annotation**] (**Figure 1F**): Gene prediction is performed using Prodigal (Hyatt et al., 2010), and genes of interest are detected using HMM models publicly available (Albertsen et al., 2013; Guglielmini et al., 2014; Grazziotin et al., 2017). However, this step is independent from the others, and any annotation tool can be applied instead.

MetaTOR generates a set of annotated metagenomics bins and their corresponding FASTA sequences (in addition to the contact network) (**Figure 1E**).

Construction of meta3C Libraries and Generation of Metagenome Assemblies

To validate and compare the pipeline to classical metagenomic binning algorithms, we investigated the gut microbiota of various mice using meta3C libraries. Feces were sampled from three groups of two mice from the Institut Pasteur animal facility, over 20 days (Materials and Methods) (**Supplementary Figure 1**). Twenty meta3C libraries (three from cage n°1, nine from cage n°2, and eight from cage n°3) were then generated as described (Marbouty et al., 2017) (*Materials and Methods*) using HpaII as restriction enzyme. Libraries were sequenced using PE Illumina 2x75 bp Kits (**Table 1**) (NCBI BioProject PRJNA542645). After trimming and quality filtering, between 25 and 100 million PE reads were recovered for each of the samples (~813 million PE reads total).

Meta3C sequences can be directly used to generate a *de novo* assembly without notable increase of false/chimeric contigs (Marbouty et al., 2014). Three assemblies (1, 2, and 3) using reads collected from cage 3/day 2, cage 3/all samples, and all cages/all samples, respectively, were generated using MEGAHIT (Li et al., 2015) (*Materials and Methods*). After discarding contigs under 500 bp, the three assemblies resulted in 61,600, 167,810, and 237,868 contigs for a cumulated size of 146, 475, and 763 Mb, respectively (**Table 2**). These assemblies and their corresponding set of reads were used to test the binning pipelines MetaTOR, MetaBAT, and CONCOCT, and their output (*Material and Methods*). The number of species present in the total assembly (n°3) was estimated using MG Rast and the alpha diversity provided for the assembly (Meyer et al., 2008) (*Material and Methods*). In total, 268 bacterial genomes are predicted to be present in the global assembly.

TABLE 1 | Meta3C libraries constructed and sequenced.

Sample	Raw paired-end reads
Cage1-day1	79 868 626
Cage1-day2	38 728 350
Cage1-day3	33 173 429
Cage2-day1	40 380 356
Cage2-day2	62 424 123
Cage2-day3	31 436 086
Cage2-day4	34 124 320
Cage2-day5	48 472 570
Cage2-day6	36 129 310
Cage2-day7	32 608 370
Cage2-day8	43 473 731
Cage2-day9	67 768 796
Cage3-day1	108 114 353
Cage3-day2	39 719 377
Cage3-day3	37 792 067
Cage3-day4	36 805 550
Cage3-day5	34 529 306
Cage3-day6	59 092 136
Cage3-day7	28 833 461
Cage3-day8	30 521 091

TABLE 2 | Assembly metrics. Only the metrics concerning assemblies filtered for the contigs above 500 bp are shown.

	PE reads (filtered)	Total size (contigs > 500 bp)	Contigs > 500 bp	N50 (contigs > 500 bp)
Assembly #1 (cage 3—day 2)	100,258,683	146,319,508 bp	61,666	6,176 bp
Assembly #2 (cage 3—samples x 8)	330,324,521	475,681,220 bp	167,810	7,578 bp
Assembly #3 (samples x 20)	813,376,239	763,455,888 bp	237,868	12,339 bp

Binning of Metagenomes Using MetaTOR

Pairs of meta3C reads were aligned independently on their respective assembly to identify those for which both reads aligned on different contigs (parameters: MQT = 20; contig size limit = 500 bp). Normalized contact scores between contigs where computed by dividing the number of pairs bridging two contigs by the square root of the product of each contig coverage. For each assembly, this step generates a network of weighted connections between contigs (**Table 3**). Each network was subsequently partitioned into CCs through iterative Louvain partitioning. After ~100 cycles, the number of large CCs (>500 kb) reaches a plateau for the three networks (**Figure 2A**). Contacts between CCs appear low, suggesting that contigs interacting preferentially with each other's were successfully pooled together (**Figure 2B**).

We analyzed, using CheckM (Parks et al., 2015), the gene content of the 17, 33, and 125 CCs > 500 kb from assemblies 1, 2, and 3, respectively. Most CCs showed completion and contamination levels above 80% and under 10%, respectively (**Figure 2C**), suggesting that they contain near-complete bacterial genomes. Those CCs were annotated as valid bins or MAGs. However, a subset of CCs displayed high contamination rate, from 10% to more than 1,000% while showing high 70/80% completion levels as well (4, 24, and 25 CCs for assemblies 1, 2, and 3, respectively) (**Figure 2C**). We suspected that these high contamination rates reflected the pooling of DNA contigs belonging to related species sharing conserved/similar sequences. We therefore applied on these CCs an extra recursive procedure consisting of processing them with 10 Louvain clustering steps. This generated sub-networks or sub-CCs (**Figure 2D**) that often display high-quality signatures of bacterial genomes, showing that indeed the large, contaminated CCs correspond to mixes of near-complete bacterial genomes (**Figure 2F**). These sub-CCs also often belonged to the same taxonomic group, suggesting that indeed sequence homology between closely related species bridged these contigs together. A focus on assembly #3 shows that the computation generated 1,001 bins > 10 kb corresponding to 724 Mb, among which 686 Mb (95%), was included within 271 bins larger than 500 kb (**Figure 2E**). This number can be compared to the 268 genomes predicted to be present in the assembly (above; *Materials and Methods*). The average completion and contamination levels of these CCs are 65.8% and 2.4%, respectively (to compare with 88.4% and 61.4% if the recursive procedure was not applied). MAG evaluation was performed (Bowers et al., 2017), resulting in 82 high-quality (< 5% contamination, > 90% completion and presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs), 87 medium-quality (< 10% contamination and > = 50% completion), and 96 low-quality MAGs (< 10% contamination and < 50% completion) (**Table 4**) (other MAGs display more than 10% of contamination; **Supplementary Table 1**).

Comparison With Hybrid Binning Algorithms

To evaluate how MetaTOR compares to existing binning approaches, we ran MetaBAT (v.1; Kang et al., 2015) and CONCOCT (Alneberg et al., 2014) on assemblies #1, #2, and #3 using the same filtered PE reads, allowing each pipeline to take advantage of the information from differential coverage across the independent experiments. The metric used to assess the efficiency of the three programs is their CheckM output (i.e., levels of completion and contamination) and the number of high-/medium-/low-quality MAGs (**Figure 3** and **Table 3**). For the three assemblies, MetaTOR retrieved 9, 41, and 82 high-quality MAGs, compared to 0, 3, and 22 with MetaBAT and 0, 11, and 12 with CONCOCT. MetaTOR also retrieved more bins exhibiting a high completion/low completion rate (90–10%) (**Figure 3**). The mean completion and contamination rates of bins characterized by MetaBAT using the 20 libraries were slightly better (respectively, 74% and 1.7%) than the ones obtained using MetaTOR (respectively, 65.8% and 2.4%) (**Figure 3**), but this could be due to the greater number of bins (>500 kb) obtained using MetaTOR (MetaBAT = 172; MetaTOR = 271) (**Table 4**). To compare further the output of MetaTOR and MetaBAT and their ability to reconstruct genomes from different phyla, we analyzed the taxonomic annotations of assembly #3 with the taxonomy of all the bins above 500 Kb retrieved for this assembly (**Supplementary Figure 2**). The bins generated by both softwares were highly

TABLE 3 | Network features.

	PE reads (filtered)	Mapped PE reads	Intercontig interactions	Weighted interactions
Assembly #1	100,258,683	67,994,798	6,457,842	1,322,003
Assembly #2	330,324,521	215,768,714	30,206,795	8,505,609
Assembly #3	813,376,239	541,384,131	96,546,376	77,577,924

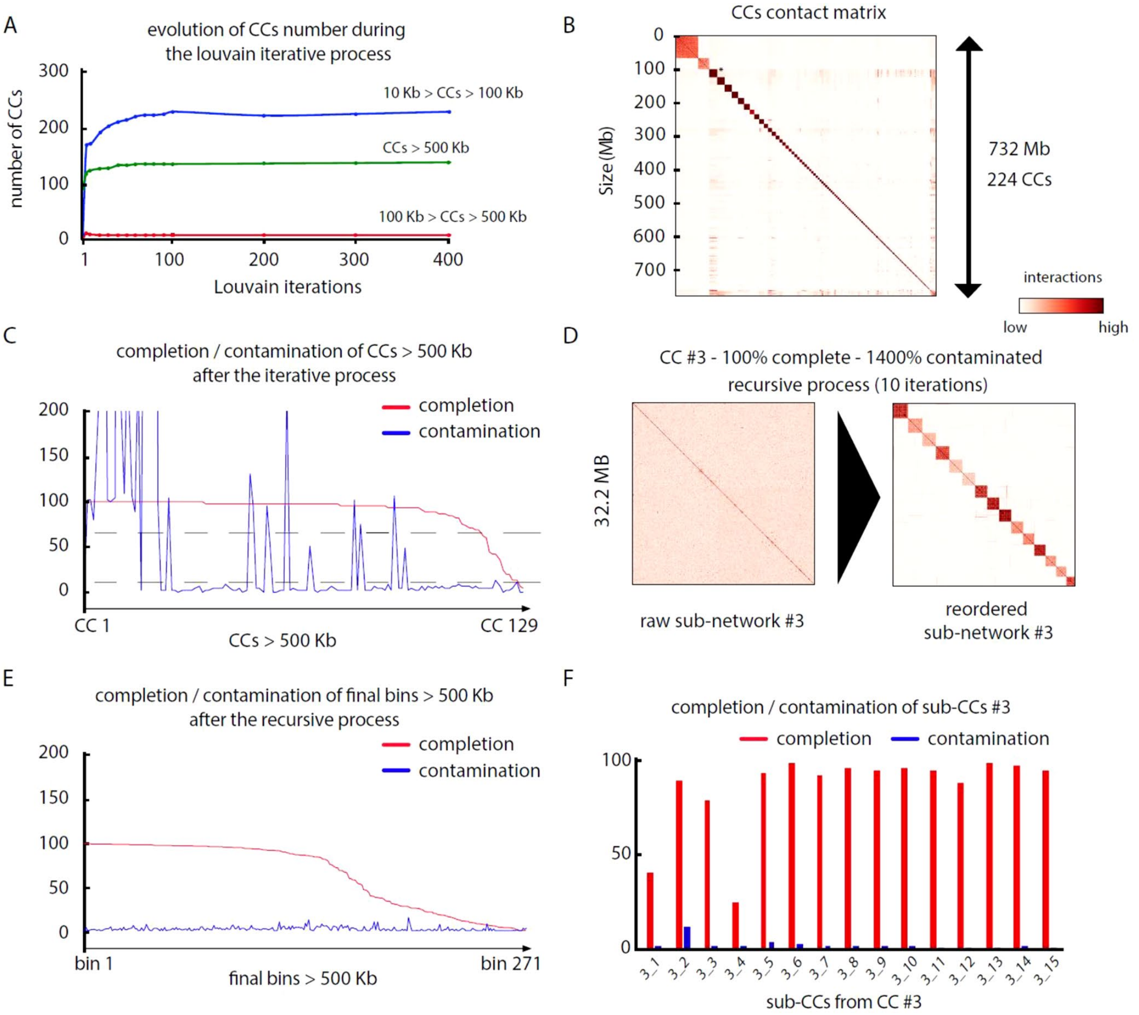


FIGURE 2 | MetaTOR partitioning of a complex microbial community. **(A)** Evolution of the number of CCs, ordered by size categories, during 400 Louvain iterations for assembly n°3 (20 samples). Color represents the amount of DNA in a given CC. Blue: 10 to 100 kb. Red: 100 to 500 kb. Green: > 500 kb. **(B)** Contact matrix encompassing the 224 largest CCs ordered by size, after 100 Louvain iterations (1 pixel = 200 kb). Y-axis: cumulated DNA size. **(C)** Completion (red) and contamination (blue) of the 129 CCs containing more than 500 kb after 100 Louvain iterations. Dashed lines: thresholds used to process CCs through a recursive procedure (completion threshold: upper 70%; contamination threshold: upper 10%). **(D)** Contact map of a highly contaminated CC (CC #3—100% complete—1,400% contaminated) before (left) and after (right) the recursive procedure (10 iterations; 1 pixel: 20 kb). Left map: contigs are ordered by size. Right map: sub-CCs are ordered by size. **(E)** Completion and contamination of the 269 CCs and sub-CCs bigger than 500 kb defined after the whole procedure. Red: completion. Blue: contamination. **(F)** Completion (red) and contamination (blue) levels of the sub-CCs retrieved from the original CC #3 after recursive procedure (10 iterations).

consistent with the assembly annotation suggesting that they do not present particular taxonomic bias in their binning process. To evaluate MAGs, 16S and 23S rRNA were searched in assembly #3 using METAXA2 (Bengtsson-Palme et al., 2015). A total of 507 23S rNRA and 304 16S rRNA were found but only 213 and 143, respectively, were located on contigs longer than 1 kb. As CONCOCT and MetaBAT only use contigs upper 1 kb, this severely decrease the amount of potential rRNA found in their bins and could explain why they were only able to bin very few high-quality drafts according to MiMAG standards (rRNA were very often the limiting factor to classify bins in that category) (Bowers et al., 2017). We then wonder if our method, which can bin contigs regardless of their size, shows better results concerning low-covered and/or highly fragmented genomes. We looked at the relation between completion for bins with a contamination rate lower than 10% and assembly statistics for those bins (**Figure 4**). Whereas we could not see clear differences between MetaBAT and MetaTOR when we look at the bins' mean coverage (**Figure 4B–D**), it appears clearly that the contigs' fragmentation is a limiting factor for MetaBAT as observed when we plotted the completion rate in function of the N50 (**Figure 4A–C**). These observations suggest that MetaTOR is able to accurately bin relatively fragmented genomes and correctly assign contigs smaller than 1 kb.

TABLE 4 | Comparison of MetaTOR, CONCOCT, and MetaBAT results.

		Assembly #1 (148 Mb)		**Assembly #2 (483 Mb)**		**Assembly #3 (763 Mb)**	
		Nb	**Size (bp)**	**Nb**	**Size (bp)**	**Nb**	**Size (bp)**
Metator	10 kb < bins < 100 kb	284	7,537,821	807	21,139,528	617	15,175,457
	100 kb < bins < 500 kb	43	11,319,827	144	30,749,287	106	22,963,515
	Bins > 500 kb	56	119,111,306	183	399,972,204	271	685,955,810
	Low-quality MAGs	31	36,042,593	97	107,071,523	96	128,486,895
	Medium-quality MAGs	16	47,397,754	39	131,055,387	87	285,670,443
	High-quality MAGs	9	35,670,959	41	140,967,746	82	259,541,396
MetaBAT	10 kb < bins < 100 kb	0	0	0	0	0	0
	100 kb < bins < 500 kb	18	5,703,905	55	17,583,986	65	24,087,225
	Bins > 500 kb	36	82,290,484	126	284,973,235	172	420,081,339
	Low-quality MAGs	14	12,478,196	44	52,797,176	95	36,277,628
	Medium-quality MAGs	21	61,439,633	73	202,719,703	143	322,230,178
	High-quality MAGs	0	0	3	5,488,345	22	58,276,800
CONCOCT	10 kb < bins < 100 kb	11	432,808	25	1,040,872	24	1,122,733
	100 kb < bins < 500 kb	7	1,351,308	23	6,275,583	6	5,193,580
	Bins > 500 kb	29	120,778,514	126	412,598,588	195	673,338,423
	Low-quality MAGs	8	17,152,380	41	76,579,222	42	70,748,222
	Medium-quality MAGs	11	25,303,368	49	134,612,509	114	358,231,099
	High-quality MAGs	0	0	11	49,146,272	12	47,807,957

DISCUSSION

We previously showed that a blind analysis of meta3C/proximity-ligation reads generated from controlled and unknown, complex mixes of species could be used to characterize efficiently their genomes (Marbouty et al., 2014; Marbouty et al., 2017). In the present work, we extend our original approach by developing a scalable computational pipeline, MetaTOR, and applying it on multiple samples of meta3C gut microbiota libraries. Compared to binning programs MetaBAT and CONCOCT, MetaTOR was able to retrieve more high-quality MAGs from highly fragmented assemblies. This work shows that physical collisions between DNA sequences represent an objective, quantitative measure to cluster these molecules together. This approach could therefore nicely complement or replace popular approaches that exploit sequence composition correlations or abundance co-variation. This remains true even when 20 independent experiments were used, highlighting the interest to include room for some meta3C experiments in future metagenomics projects, regardless of the number of planed libraries. Meta3C-like methods have only been applied to microbial rich samples so far (mice and human gut, cow rumen, sewage) (Marbouty et al., 2017; Stewart et al., 2018; Stalder et al., 2019) and still need to be improved in order to be applied to sample with low concentration of microorganisms. The time needed to generate a meta3C library is 3 days, and up to 16 libraries can be generated in parallel (Foutel-Rodier et al., 2018). It is also likely that commercial kits will be available relatively soon, which will boost the amenability of the approach. The cost of a single library is estimated to ~50€ (not including processing and sequencing of the library). Therefore, we believe this approach is well suited for a variety of metagenomics projects.

A limitation of the present work consists in the size of the reads sequences, 65 bp, whereas most metagenomics studies sequence longer reads (150 bp). This is probably a disadvantage for the two binning programs we tested as the assembly quality is technically lower than what it would have been if computed with longer reads. On the other hand, one could also argue that meta3C/MetaTOR can therefore be performed using cheaper, short-read sequencing approaches and still provide good results. However, more tests are needed to fully characterize the influence of assembly quality on the different programs in light of MetaTOR results.

To improve the assembly, regardless of the read length, it is also possible to apply the approach used in Marbouty et al. (2017), which consists in mapping back the total reads (including ambiguous ones originally discarded) back to contigs of one bin. These reads are then used to generate a new assembly for each individual bin. Although time consuming, we showed that this approach improved the assembly statistics of each bin (Marbouty et al., 2017).

The large networks derived from different meta3C libraries contain several highly connected sub-networks poorly connected to each other. Highly modular networks such as those are known to be well-suited for community detection algorithm like Louvain (Blondel et al., 2008). Moreover, the "iterative Louvain" procedure allows us to identify sets of sequences that contact each other. However, there are limits to the current iterative Louvain implementation. First, all modularity optimization algorithms tend to over-cluster nodes when the network reaches a certain size threshold, regardless of the underlying patterns. This well-documented property is known as the "resolution limit" (Fortunato and Barthélemy, 2007). However, it can be sidestepped by running the partitioning process recursively on the network corresponding to the studied sub-network. Since it should be comparatively small and under the scale at which the aforementioned limit becomes visible, the clusters found inside will separate again and yield bins as normal. The recursive procedure applied in the present work appears as highly effective

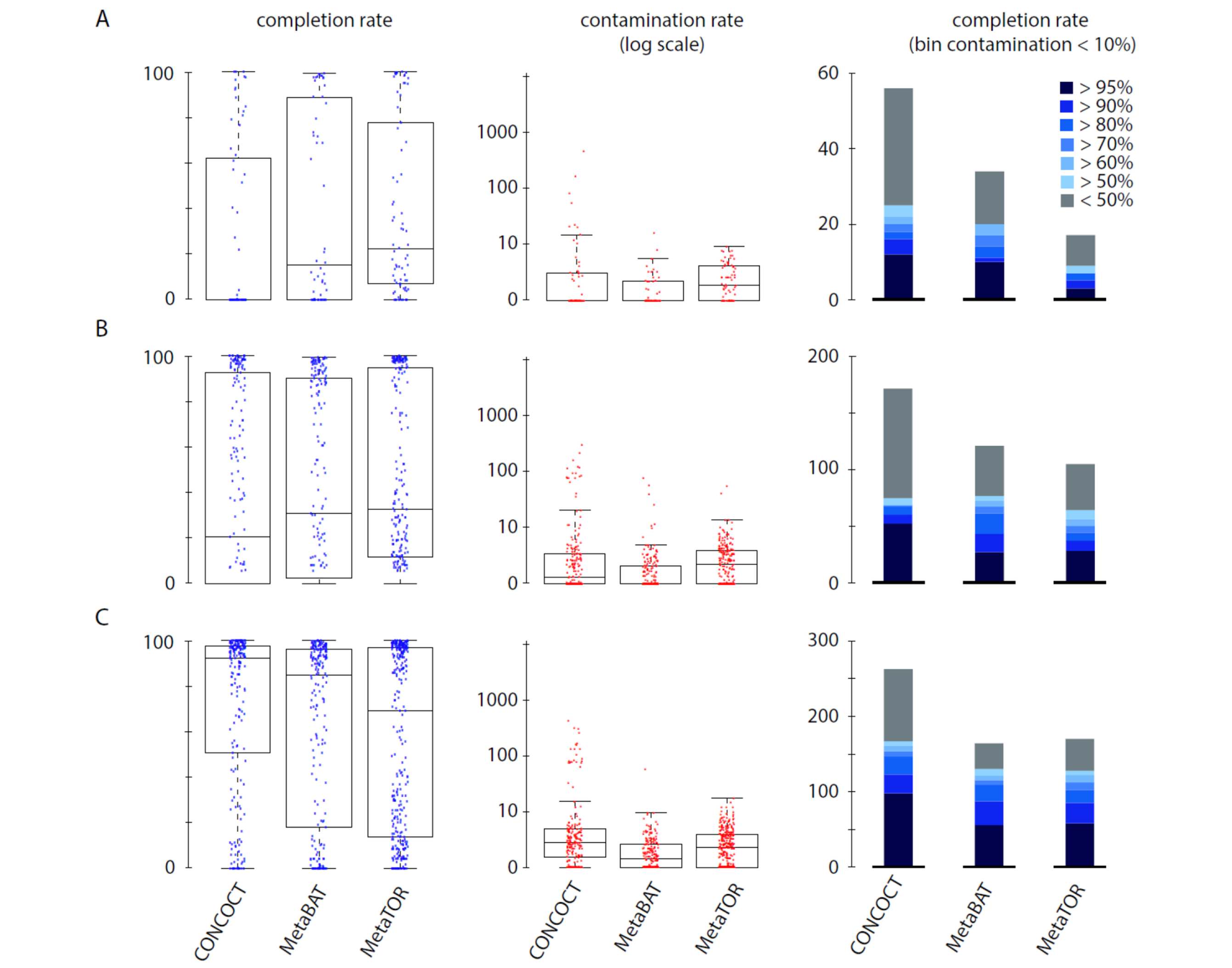


FIGURE 3 | Comparison of MetaTOR, MetaBAT, and CONCOCT. CheckM output comparison for the three binning methods applied on the three assemblies tested in this work. **(A)** Assembly 1 (one meta3C library). **(B)** Assembly 2 (eight libraries). **(C)** Assembly 3 (20 libraries). Box plot for completion (left) and contamination (middle) and histogram of retrieved MAGs (right) are presented for the three binning methods. Only MAGs over 500 kb and harboring less than 10% of contamination are analyzed.

with a clear increase in the number of high-quality MAGs retrieved.

A second limit comes from the stringent definition of CCs that consist of sequences that always, systematically cluster together. As a result, a single "jump" of a contig out of a cluster during one of the iterations will lead to its exclusion from the final CCs. While this allows contamination reduction, a number of meaningful sequences could still incidentally be excluded from the bin. Indeed, mobile or repeated elements (e.g., phage, prophages, or plasmids) shared by different species will likely be excluded from their corresponding CCs. This limitation can be overcome *a posteriori* as follows. First, annotation pipelines such as VirSorter (Roux et al., 2015) or PlasFlow (Krawczyk et al., 2018) allows to identify contigs carrying such sequences. Second, the bacterial hosts of these contigs can be inferred using the contact network as described in (Marbouty et al., 2017), and/or with the help of the Louvain clustering score (computed from the iterative procedure, and corresponding to the number of times two CCs are grouped together). A detailed analysis of overlapping communities would be very useful in the future to study such associations and bring a new tool in the study of interactions between genomic entities in microbial communities.

Our pipeline is flexible: although it was developed to take advantage of the Louvain algorithm, other clustering algorithms yielding nondeterministic community identifiers (e.g., a community detection algorithm with a different modularity) can be used instead with no side effects on the rest of the pipeline.

Proximity-ligation assays were originally developed to capture the 3D folding of microbial or mammalian chromosomes (Dekker et al., 2002; Lieberman-Aiden et al., 2009). Derivative applications of these techniques were

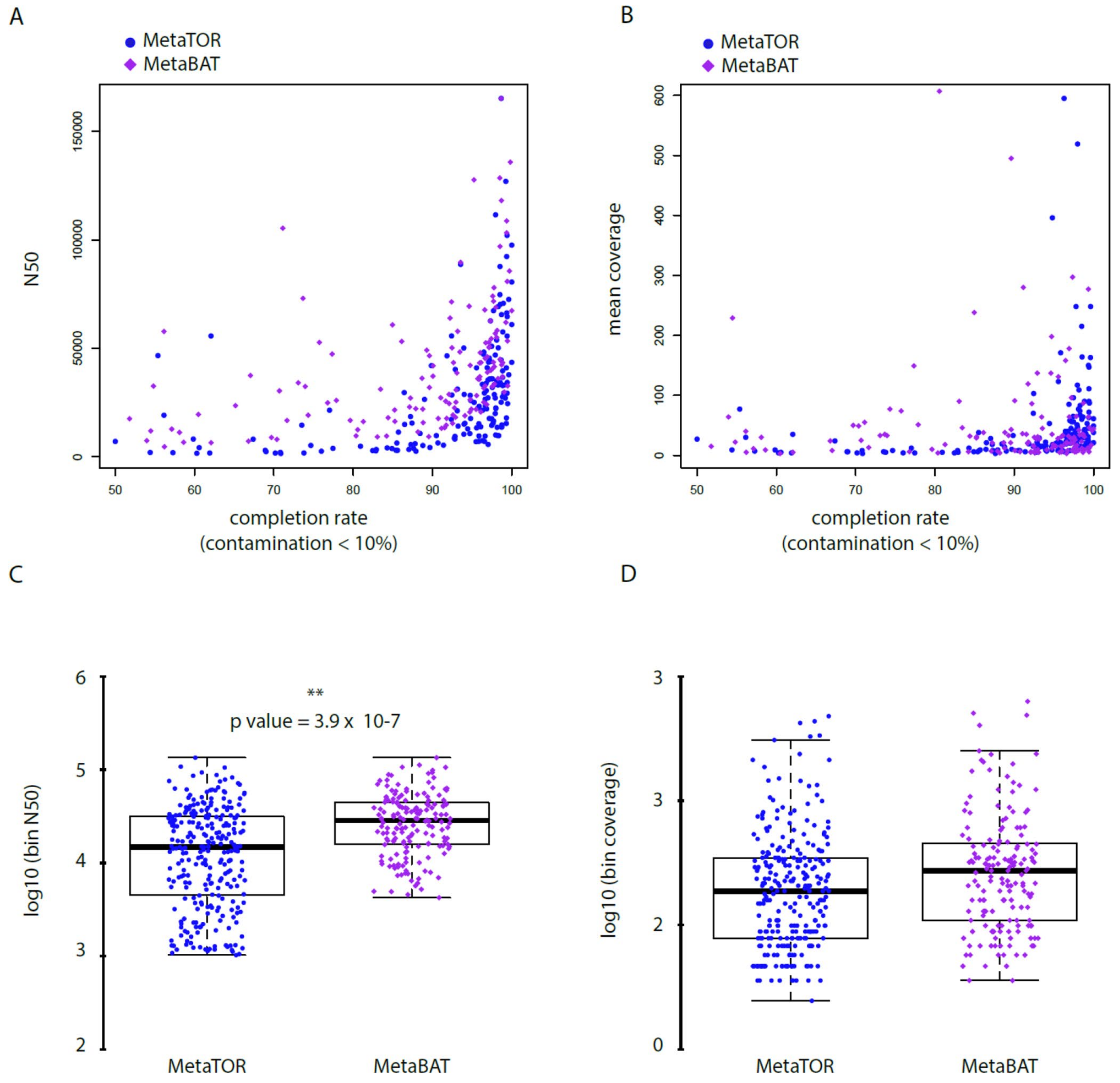


FIGURE 4 | Statistics of low contaminated reconstructed bins. **(A–B)** Correlation between completion rate and N50 **(A)** or mean coverage **(B)** for bins with a contamination rate below 10%. Blue circles = MetaTOR bins. Purple diamonds = MetaBAT bins.) **(C–D)** Box plot for N50 **(C)** and mean coverage **(D)** of retrieved bins with a contamination rate below 10% are presented for MetaTOR (blue circles) and MetaBAT (purple diamonds). A t-test shows a clear difference between distribution of bins' N50 for the two software (**C**—p-value = 3.9 x 10^{-7}).

developed and applied to solve or improve genomics techniques such as chromosome-level scaffolding (Kaplan and Dekker, 2013; Burton et al., 2013; Marie-Nelly et al., 2014), haplotype reconstruction (Selvaraj et al., 2013), or centromere annotation (Marie-Nelly et al., 2014). Haplotype phasing is a particularly interesting development to combine with metagenomics data since strains from the same species remain challenging to characterize. This requires both an improvement in meta3C like capture yield to increase the resolution in coverage of the contigs, as well as the integration of computational haplotype phasing programs.

DATA AVAILABILITY

The datasets generated for this study can be found on SRA database: BioProject PRJNA542645.

ETHICS STATEMENT

Animal experimentation: The Institut Pasteur ethics organism (CETEA) approved all the experiments performed on mice (Project dha170005).

AUTHOR CONTRIBUTIONS

MM and RK conceived the study. LB, TFR and MM wrote the pipeline MetaTOR. MM, TFR, and AT performed the experiments. LB, TFR, MM, and RK analyzed and interpreted the data. LB, TFR, MM, and RK wrote the manuscript.

FUNDING

LB is supported by an AMX fellowship from the French Ministry of Higher Education, Research and Innovation. TFR is supported by an ENS fellowship by the French Ministry of Higher Education, Research and Innovation. This research was supported by funding to RK from the European Research Council under the Horizon 2020 Program (ERC grant agreement 260822) and from the Agence Nationale pour la Recherche (JPI-EC-AMR STARCS ANR-16-JPEC-0003-05).

ACKNOWLEDGMENTS

We thank Corinne Fayolle and Xavier Montagutelli for their help in the sampling process. We thank our colleagues from the lab for discussions, feedback, and comments on MetaTOR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00753/full#supplementary-material

REFERENCES

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31 (6), 533–538. doi: 10.1038/nbt.2579

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 (11), 1144–1146. doi: 10.1038/nmeth.3103

Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., et al. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing Proximity Ligation Products. *PeerJ* 2, e415. doi: 10.7717/peerj.415

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., et al. (2015). METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* 15 (6), 1403–1414. doi: 10.1111/1755-0998.12399

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory E* (10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35 (8), 725–731. doi: 10.1038/nbt.3893

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125.

Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014). Species-level deconvolution of metagenome assemblies with Hi-C–based contact probability maps. *G3 (Bethesda)* 4 (7), 1339–1346. doi: 10.1534/g3.114.011825

Castelle, C. J., and Banfield, J. F. (2018). Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* 172 (6), 1181–1197. doi: 10.1016/j.cell.2018.02.016

Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* 13 (4), 260–270. doi: 10.1038/nrg3182

Coutinho, F. H., Gregoracci, G. B., Walter, J. M., Thompson, C. C., and Thompson, F. L. (2018). Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends Microbiol.* 26 (11), 955–965. doi: 10.1016/j.tim.2018.05.015

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295 (5558), 1306–1311. doi: 10.1126/science.1067799

DeMaere, M. Z., and Darling, A. E. (2019). Bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Bio.* 20 (1), 46. doi: 10.1186/s13059-019-1643-1

Edbeib, M. F., Wahab, R. A., and Huyop, F. (2016). Halophiles: biology, adaptation, and their role in decontamination of hypersaline environments. *World J. Microbiol. Biotechnol.* 32 (8), 135. doi: 10.1007/s11274-016-2081-9

Flot, J.-F., Marie-Nelly, H., and Koszul, R. (2015). Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. *FEBS Lett.* 589 (20 Pt A), 2966–2974. doi: 10.1016/j.febslet.2015.04.034

Fortunato, S., and Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci.* 104 (1), 36–41. doi: 10.1073/pnas.0605965104

Foutel-Rodier, T., Thierry, A., Koszul, R., and Marbouty, M. (2018). Generation of a metagenomics proximity ligation 3C library of a mammalian gut microbiota. *Methods Enzymol.* 612, 183–195. doi: 10.1016/bs.mie.2018.08.001

Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G. H., McHardy, A. C., Nederbragt, A. J., et al. (2016). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.* 6, 25373. doi: 10.1038/srep25373

Graham, E. D., Heidelberg, J. F., and Tully, B. J. (2017). BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 5, e3035. doi: 10.7717/peerj.3035

Grazziotin, A. L., Koonin, E. V., and Kristensen, D. M. (2017). Prokaryotic virus orthologous groups (PVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 45 (D1), D491–DD98. doi: 10.1093/nar/gkw975

Guglielmini, J., Néron, B., Abby, S. S., Garcillán-Barcia, M. P., de la Cruz, F., and Rocha, E. P. C. (2014). Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* 42 (9), 5715–5727. doi: 10.1093/nar/gku194

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1 (5), 16048. doi: 10.1038/nmicrobiol.2016.48

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

Ji, P., Zhang, Y., Wang, J., and Zhao, F. (2017). MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat. Commun.* 8, 14306. doi: 10.1038/ncomms14306

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165. doi: 10.7717/peerj.1165

Kaplan, N., and Dekker, J. (2013). High-Throughput Genome Scaffolding from in-Vivo DNA Interaction Frequency. *Nat. Biotechnol.* 31, (12) 1143–1147. doi. 10.1038/nbt.2768

Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* 46, e35. doi: 10.1093/nar/gkx1321

Laczny, C. C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2017). BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.* 45 (W1), W171–W179. doi: 10.1093/nar/gkx348

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35 (9), 3100–3108. doi: 10.1093/nar/gkm160

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 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950), 289–293. doi: 10.1126/science.1181369

Long, P. E., Williams, K. H., Hubbard, S. S., and Banfield, J. F. (2016). Microbial metagenomics reveals climate-relevant subsurface biogeochemical processes. *Trends Microbiol.* 24 (8), 600–610. doi: 10.1016/j.tim.2016.04.006

Lowe, T. M., and Eddy, S. R. (1997). TRNAscan-SE: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res.* 25 (5), 955–964. doi: 10.1093/nar/25.5.0955

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 (6), 791–798. doi: 10.1093/bioinformatics/btw290

Marbouty, M., Baudry, L., Cournac, A., and Koszul, R. (2017). Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* 3 (2). doi: 10.1126/sciadv.1602105

Marbouty, M., Cournac, A., Flot, J. F., Nelly, H. M., Mozziconacci, J., and Koszul, R. (2014). Metagenomic chromosome conformation capture (Meta3C) unveils the diversity of chromosome organization in microorganisms. *ELife* 3, e03318. doi: 10.7554/eLife.03318

Marbouty, M., and Koszul, R. (2015). Metagenome analysis exploiting high-throughput chromosome conformation capture (3C) data. *Trends Genet.* 31 (12), 673–682. doi: 10.1016/j.tig.2015.10.003

Marbouty, Ma., Le Gall, A., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J. B., et al. (2015). Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Mol. Cell* 59 (4), 588–602. doi: 10.1016/j.molcel.2015.07.020

Marie-Nelly, H., Marbouty, M., Cournac, A., Liti, G., Fischer, G., Zimmer, C., et al. (2014). Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics* 30 (15), 2105–2113. doi: 10.1093/bioinformatics/btu162

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17 (1), 10–12. doi: 10.14806/ej.17.1.200

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al., (2008). The Metagenomics RAST Server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386. doi: 10.1186/1471-2105-9-386

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., et al. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinformatics*. doi: 10.1093/bib/bbx098

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., et al. (2016). Uncovering earth's virome. *Nature* 536 (7617), 425–430. doi: 10.1038/nature19094

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25 (7), 1043–1055. doi: 10.1101/gr.186072.114

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and van der Putten, W. H. (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* 11 (11), 789–799. doi: 10.1038/nrmicro3109

Press, M. O., Wiser, A. H., Kronenberg, Z. N., Langford, K. W., Shakya, M., Lo, C.-C., et al. (2017). Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *BioRxiv*, 198713. doi: 10.1101/198713

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35 (9), 833–844. doi: 10.1038/nbt.3935

Rosado, P. M., Leite, D. C. A., Duarte, G. A. S., Chaloub, R. M., Jospin, G., Nunes da Rocha, U., et al. (2019). Marine probiotics: increasing coral resistance to bleaching through microbiome manipulation. *ISME J.* 13 (4), 921. doi: 10.1038/s41396-018-0323-6

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3, e985. doi: 10.7717/peerj.985

Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* 31 (12), 1111–1118. doi: 10.1038/nbt.2728

Sieber, C. M. K., Probst, A. J., Sharrar, A., BThomas, C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes *via a* dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3 (7), 836. doi: 10.1038/s41564-018-0171-1

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., et al. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521 (7551), 173–179. doi: 10.1038/nature14447

Stalder, T., Press, M. O., Sullivan, S., Liachko, I., and Top, E. M. (2019). Linking the resistome and plasmidome to the microbiome. *ISME J.* 1–10. doi: 10.1038/s41396-019-0446-4

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* 9 (1), 870. doi: 10.1038/s41467-018-03317-6

Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2 (1), 26. doi: 10.1186/2049-2618-2-26

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

frontiers in Genetics

TECHNOLOGY AND CODE
published: 13 September 2019
doi: 10.3389/fgene.2019.00849

"EviMass": A Literature Evidence-Based Miner for Human Microbial Associations

Divyanshu Srivastava[1†], *Krishanu D. Baksi*[1,2†], *Bhusan K. Kuntal*[1,3,4*] and *Sharmila S. Mande*[1*]

[1] Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune, India, [2] School of Information Technology, Indian Institute of Technology Delhi, Delhi, India, [3] Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune, India, [4] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

OPEN ACCESS

Edited by:
Bas E. Dutilh,
Utrecht University, Netherlands

Reviewed by:
Jia Qu,
China University of Mining and Technology, China
Richard R. Rodrigues,
Oregon State University, United States
Basten Snoek,
Utrecht University, Netherlands

***Correspondence:**
Bhusan K. Kuntal
kuntal.bhusan@tcs.com
Sharmila S. Mande
sharmila.mande@tcs.com

†These authors have contributed equally to this work

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

Received: *24 April 2019*
Accepted: *14 August 2019*
Published: *13 September 2019*

Citation:
Srivastava D, Baksi KD, Kuntal BK and Mande SS (2019) "EviMass": A Literature Evidence-Based Miner for Human Microbial Associations. Front. Genet. 10:849. doi: 10.3389/fgene.2019.00849

The importance of understanding microbe–microbe as well as microbe–disease associations is one of the key thrust areas in human microbiome research. High-throughput metagenomic and transcriptomic projects have fueled discovery of a number of new microbial associations. Consequently, a plethora of information is being added routinely to biomedical literature, thereby contributing toward enhancing our knowledge on microbial associations. In this communication, we present a tool called "EviMass" (Evidence based mining of human Microbial Associations), which can assist biologists to validate their predicted hypotheses from new microbiome studies. Users can interactively query the processed back-end database for microbe–microbe and disease–microbe associations. The EviMass tool can also be used to upload microbial association networks generated from a human "disease–control" microbiome study and validate the associations from biomedical literature. Additionally, a list of differentially abundant microbes for the corresponding disease can be queried in the tool for reported evidences. The results are presented as graphical plots, tabulated summary, and other evidence statistics. EviMass is a comprehensive platform and is expected to enable microbiome researchers not only in mining microbial associations, but also enriching a new research hypothesis. The tool is available free for academic use at https://web.rniapps.net/evimass.

Keywords: microbiome, literature mining, human disease, web server, microbial association

INTRODUCTION

The microbial groups residing in human body remain in complex association within themselves as well as with the host. These associations range from mutualism, amenalism, and commensalism to parasitism, predation, and competitions (Faust and Raes, 2012). However, with the onset of a disease, the human microbiome is often seen to display aberrations, which may be a cause or an effect (Eloe-Fadrosh and Rasko, 2013; Liang et al., 2018). Advances in the field of metagenomics have made it possible to successfully capture and report such microbial dysbiosis observed in the diseased state. Microbial abundance measurements for many samples can be simultaneously obtained using 16S rRNA (amplicon) sequencing in a short span of time (Goodrich et al., 2014). Recent developments in sequencing technology and the drastic reduction in the associated cost have encouraged researchers to probe the microbial basis of various human diseases. Consequently, a plethora of information relating to microbes and their association with diseases are added to the growing biomedical literature (Cani, 2018). Although the obtained microbiome data can be used to calculate differentially abundant genera as well as their co-occurrence patterns (Kuntal et al., 2013; Kumar et al., 2014; Dhariwal et al., 2017), their evidence from biomedical literature can help to strengthen a research hypothesis.

The Human Microbe Disease Association Database (HMDAD) was the first resource developed using literature mining to systematically gather experimental data to study microbe–disease associations (Ma et al., 2017b). Several tools have been developed thereafter to utilize the curated data from HMDAD and score human microbe associations using advanced mathematical approaches (Chen et al., 2017; Huang et al., 2017a; Huang et al., 2017b; Wang et al., 2017b; Peng et al., 2018; Zou et al., 2018; Qu et al., 2019). The above set of tools focuses on identifying associated genera across a set of selected diseases and is eventually used to find diseases having similar pattern of associated microbes. For example, KATZHMDA (Chen et al., 2017) computes the number of walks of connections between microbe and disease nodes, LRLSHMDA (Wang et al., 2017b) uses a semisupervised learning framework based on Laplacian regularized least squares, ABHMDA (Peng et al., 2018) uses an Adaptive Boosting model, PBHMDA (Huang et al., 2017b) calculates the Gaussian interaction profile kernel similarity, and very recently a new method called MDLPHMDA (Qu et al., 2019) based on Matrix Decomposition and Label Propagation has been introduced. While some of the above methods are limited to predict microbes associated with a fixed set of diseases, more recent methods like ABHMDA can predict microbes associated with a new disease (Peng et al., 2018). In addition, methods like MDLPHMDA also can now be used to predict novel microbe–disease associations with minimum noise (Qu et al., 2019). Tools like Micro-pattern, on the other hand, can perform an enrichment analysis for a given set of microbes using a hypergeometric test (Ma et al., 2017a). This method relies on creation of pregenerated microbe sets using manual curation from selected diseases, making it advantageous for accurate predictions, but limits the applicability. Given the scenario, although the association of individual microbes with a disease can give informative predictions, the knowledge of microbial co-occurrence patterns can augment it further to provide improved insights. As microbes are known to work in mutual associations rather than single entities, it is also imperative to validate a known co-occurrence pattern observed in an experimental microbiome study. One such method called "Microbial Prior Lasso" (or MPLasso) uses literature evidence supplied as an input to quantify microbial associations and is available as an R package (Lo and Marculescu, 2017). However, the major limitation lies in gathering systematic information relating to intermicrobe association and their relation to human diseases.

In order to address the aforementioned limitation, we have developed a web-based GUI resource called "EviMass" (Evidence based mining of human Microbial Associations) available at https://web.rniapps.net/evimass that can be interactively used for not only querying microbe disease associations, but also inferring the intermicrobe association patterns mined from biomedical literature (**Figure 1**). The EviMass backend database has been developed using extensive data mining of the currently available PubMed abstracts. The front-end is designed with an interactive query system, which allows users to find all microbes associated with a user-defined query microbe. In addition, the identified microbial associations can also be visualized for their occurrence statistics in various human diseases. Similarly, users can search for an individual microbe to view all diseases associated with it and vice versa. Additionally, users can upload a microbial association network generated from experimental microbiome data corresponding to a human disease and easily verify these associations using the evidence statistics. A list of differentially abundant genera obtained from a disease–control microbiome case study can also be validated using EviMass along with an option for enrichment analysis. All evidence inferred using the present tool is listed with corresponding PubMed IDs, which can be used for further reference. The utility of EviMass is demonstrated with case studies as well as using real-world microbiome data.

RESULTS

Global Overview of Disease–Microbe Associations Captured by EviMass

EviMass backend database was generated using a systematic literature mining approach (details in *Material, Methods and Implementation*) specific to microbiome and human diseases. We focused our analysis on 51 widely reported microbiome associated human diseases and their associations with various microbes (genera level). These diseases spanned six categories, namely, systemic diseases and those affecting gut, skin, lung, brain, and urogenital system (**Table 1**). The results of the literature mining as incorporated in EviMass yielded several interesting findings. For example, ulcer, diarrhea, HIV, urinary tract infection, and cystic fibrosis were found to be the most widely (top 5) reported diseases with microbial associations (**Supplementary Figure 1**). On the other hand, microbial genera, namely, *Escherichia*, *Staphylococcus*, *Pseudomonas*, *Bacillus*, and *Streptococcus*, were seen to occupy the top 5 spots in terms of their reported all-microbiome articles in PubMed (irrespective of disease association) (**Supplementary Figure 2**). A closer look into the genera maximally associated with human diseases revealed *Escherichia*, *Lactobacillus*, *Clostridium*, *Streptococcus*, and *Bacteroides* to be the top 5 players (**Supplementary Figure 3**). A deeper analysis revealed the following genera to be significantly ($P < 0.05$) associated with diseases (affecting various organs): *Clostridium* with gut, *Staphylococcus* with skin, *Pseudomonas* with lungs, *Escherichia* with brain as well as urogenital, and *Helicobacter* with the other systemic diseases (**Figure 2**).

In order to check which all genera are closely associated with each of the aforementioned top genera irrespective of

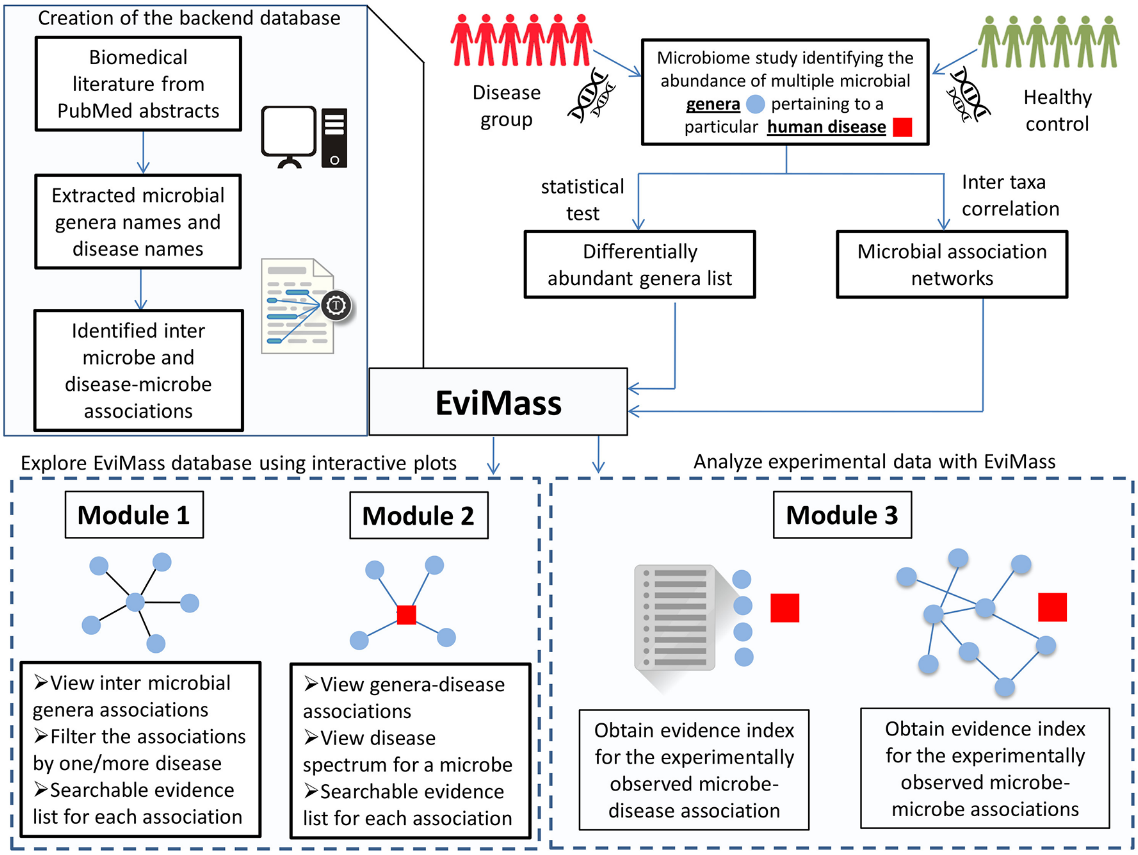

Creation of the backend database
Biomedical literature from PubMed abstracts
Extracted microbial genera names and disease names
Identified inter microbe and disease-microbe associations
Disease group
Microbiome study identifying the abundance of multiple microbial genera ● pertaining to a particular human disease 🟥
Healthy control
statistical test
Inter taxa correlation
Differentially abundant genera list
Microbial association networks
EviMass
Explore EviMass database using interactive plots
Analyze experimental data with EviMass
Module 1
Module 2
Module 3
➢View inter microbial genera associations
➢Filter the associations by one/more disease
➢Searchable evidence list for each association
➢View genera-disease associations
➢View disease spectrum for a microbe
➢Searchable evidence list for each association
Obtain evidence index for the experimentally observed microbe-disease association
Obtain evidence index for the experimentally observed microbe-microbe associations

FIGURE 1 | Overview of the EviMass backend creation and utility of its various modules in understanding the intermicrobial and microbe–disease associations.

diseases, Module 1 of the EviMass tool was utilized. The results (**Supplementary Figures 4–19**) showed a wide range of association patterns between each of these genera shown as graphs. While the central node of the graph represented the query genera, the remaining nodes corresponded to the genera associated with it. The size of the nodes depended on the strength of the associations calculated as the sum total of publications where the two genera were identified to co-occur. It was interesting to observe that most of the association graphs were dominated by a selected group of genera like *Escherichia*, *Staphylococcus*, and *Pseudomonas*. In order to get a deeper insight into the microbe–disease associations, a summary of the associated microbial genera count corresponding to each disease and the number of articles reporting the disease was generated (**Figure 3**). The Module 2 of EviMass was then used to explore each of these associations along with the literature evidences. Our analysis using EviMass for the top diseases across each category showed some amount of genera specificity (**Supplementary Figures 20–24**). For example, cystic fibrosis (**Supplementary Figure 20**) showed a very strong association with the genera *Pseudomonas* with 3,711 evidences (journal articles). Apart from being dominant in cystic fibrosis, *Pseudomonas* was also found to be associated with other diseases like HIV, diabetes, ulcer, and urinary tract infection although with lower evidences. Similar associations were also observed in other diseases (**Supplementary Figures 20–24**), which instigated an interest to look into the disease similarities based on their associated genera as explored in the next section.

Disease Similarity Based on Literature Evidence Using EviMass

Although earlier studies (Ma et al., 2017a) have shown an overall relation between various diseases based on their microbial associations, we focused on obtaining categorical insight based on our extended database (**Figure 4**). The top 20 persistent microbes across the six categories (**Table 1**) were chosen and used to generate bidirectional clustered (UPGMA hierarchical clustering) heat map for each category. Euclidean distance was used as the measure of distance, and the values were normalized by rows (diseases). Diseases like colorectal carcinoma, colon cancer, inflammatory bowel disease, irritable bowel syndrome, colitis, and kidney stones were part of closely linked cluster in the gut category. These diseases were seen to be reported with an increased association with *Lactobacillus*, *Bifidobacterium*, and *Clostridium*. The skin, brain, and urogenital diseases did not show any distinct clustering, but *Staphylococcus*, *Escherichia*, and *Lactobacillus* were observed to be the dominant players in these diseases, respectively. Asthma and related diseases were seen to cluster away from cystic

TABLE 1 | List of different microbe-related human diseases categorized by the organs they affect.

Organs affected	Diseases	No. of diseases
Gut	End-stage renal disease (ESRD), kidney stones, diarrhea, liver cirrhosis, malnutrition, ileal Crohn disease (CD), necrotizing enterocolitis, colon cancer, infectious colitis, constipation, colitis, ulcerative colitis, Whipple disease, irritable bowel syndrome (IBS), gastroesophageal reflux, Crohn disease (CD), gastric and duodenal ulcer, inflammatory bowel disease (IBD), *Clostridium difficile* infection (CDI), colorectal carcinoma	20
Skin	Skin and mucosal infections, atopic dermatitis, psoriasis, guttate psoriasis, atopic sensitization, eczema, atopy	7
Lungs	Asthma, allergic asthma, recurrent wheeze, chronic obstructive pulmonary disease, cystic fibrosis	5
Brain	Multiple sclerosis, Parkinson's disease, Schizophrenia, Autism, Depression	5
Urogenital	Urinary tract infection, bacterial vaginosis, polycystic ovary syndrome, preterm birth	4
Systemic	Type 1 diabetes, diabetes, type 2 diabetes, HIV/AIDS, obesity, systemic inflammatory response syndrome, allergic sensitization, allergy, ulcer, periodontitis	10
	Total	51

fibrosis and chronic obstructive pulmonary disease in the lung category. The remaining category of systemic diseases showed a clear cluster of allergy, obesity, and type 2 diabetes dominated by *Lactobacillus* and *Helicobacter*. Periodontitis, one of the diseases in the last category, clustered away from other systemic diseases and was characterized by the increase in association of the genera *Porphyromonas*. The "word cloud" feature was used to understand the associations that distinctly showed the dominance of the word "gingivalis" (**Supplementary Figure 25**) in the abstracts indicating the role of *Porphyromonas gingivalis*. A secondary search on the listed abstracts by using the keyword "inflammation" further yielded keywords like "cytokines," "tnf," "lps" (**Supplementary Figure 25**), which are indicators of some mechanisms of *Porphyromonas gingivalis* infection in periodontitis (Jiang et al., 2018; Kajiura et al., 2018; Zhou et al., 2018). However, these observations only provided a global picture, which can be enriched by augmenting with experimental data. In the next section, we investigated a specific disease along with reported experimental data to get more insights into microbial pathogenesis.

Case Study With Real World Microbiome Data

One of the featured utility of the EviMass tool pertains to Module 3, which allows users to validate their results from microbiome experiments based on the curated literature evidence. In order to demonstrate the utility, we first selected a publicly available data (Fazlollahi et al., 2018) where the authors studied 72 asthma subjects (using 16S ribosomal RNA sequencing on nasal swabs) and compared the same with those obtained from healthy controls. Four microbial genera reported to be significantly associated with asthma, namely, *Prevotella*, *Dialister*, *Gardnerella*, and *Alkanindiges*, were used as input for the EviMass Module 3 along with the disease keyword "asthma." The result indicated *Prevotella* to be the most widely reported as well as statistically significant ($P < 0.001$) genera to be associated, among others, for asthma (**Supplementary Figure 26**). The node "*Prevotella*" can be clicked to populate the list of PubMed articles reporting the association, which in turn can be filtered based on search criteria. As most microbes are known to orchestrate an inflammatory disease by altering the immune response in the host, we searched for the keyword "immune" to filter the articles reporting the immunological role of *Prevotella* in asthma. The search result yielded three articles, of which one clearly reported the marked capacity of *Prevotella* in driving T_H17 immune responses (Larsen, 2017).

In the next step, we used another dataset for analyzing a microbial association network for allergic asthma where the authors did not find any differentially abundant genera specific to the allergy samples (Hevia et al., 2016). We had used the same data in one of our earlier works (Kuntal et al., 2019) to identify microbial "driver" genera (using "NetShift" methodology). While *Granulicatella* and *Turicibacter* were seen to be two potential pathogenic drivers, only *Granulicatella* was predicted to be the main driver (Kuntal et al., 2019). The same microbial network was used as an input for EviMass, and the associations of *Granulicatella* and *Turicibacter* were investigated with Module 3 (also provided as an autoload example in the web server). The evidence statistics for *Granulicatella* and its associated genera (which were mostly pathogens) *Staphylococcus*, *Streptococcus*, and *Veillonella* showed a tendency to co-occur irrespective of disease condition (**Supplementary Figure 27**). For example, evidence for association of *Granulicatella* and *Staphylococcus* was seen in 23 articles, *Granulicatella* and *Streptococcus* in 80 articles, and *Granulicatella* and *Veillonella* in 35 articles. This observation provides evidence that co-occurrence of the genus *Granulicatella* with the above pathogens is indeed seen globally. On the other hand, the associations of *Turicibacter* (with *Fusibacter* and *Alkaliphilus*) did not show any literature evidence of co-occurrence (**Supplementary Figure 28**), thereby strengthening our earlier prediction of inability of *Turicibacter* to become a pathogenic driver. The primary intention of this case study was to demonstrate the ease with which scientific hypothesis in microbiome research can be enriched using the EviMass tool.

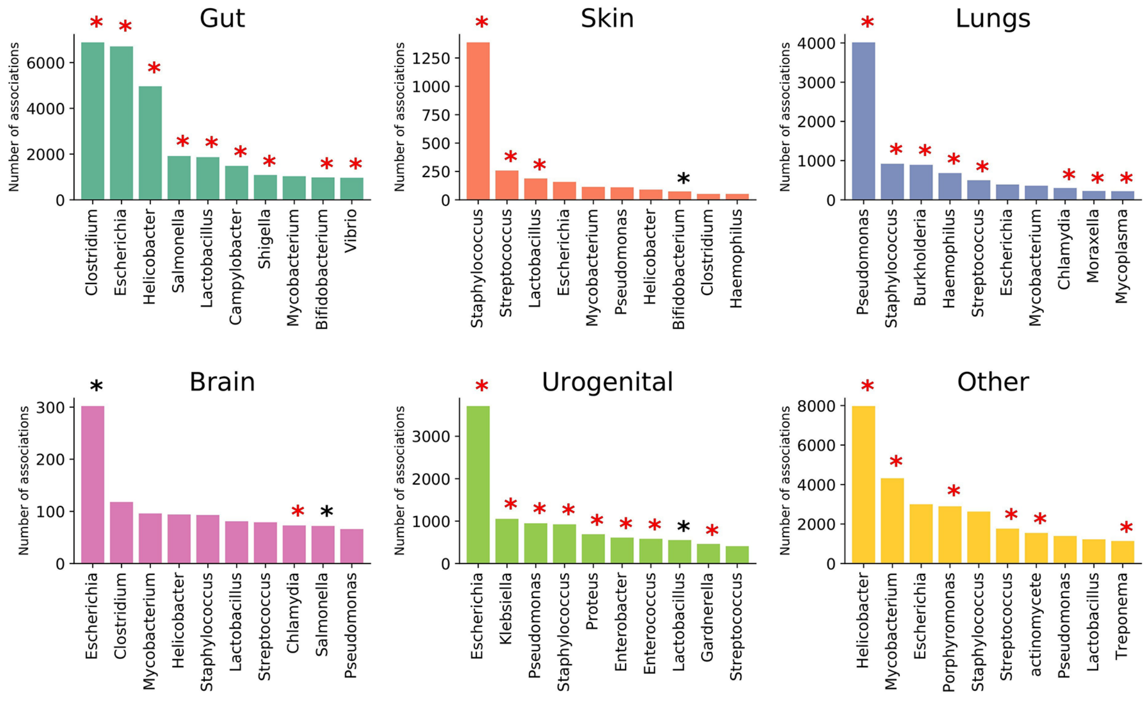


FIGURE 2 | Top 10 prominent microbial genera associated with diseases affecting various organs. Statistically significant (*P* < 0.05) genera are marked with a black asterisk (with Bonferroni-corrected *P* < 0.05 highlighted in red).

CONCLUSIONS AND FUTURE WORK

In this communication, we developed a resource for understanding the microbe–microbe and microbe–disease associations. The present version aims to provide a one-stop platform for validating data-driven hypothesis on microbiome studies. We aim to update our resource on a regular basis in order to incorporate the growing corpus of information. The current version of EviMass performs a text processing of the available PubMed abstracts to identify microbe association trends (increase or decrease). Additionally, it allows one to filter the results based on specific queries like genera/species name, journal information, or any generic keyword available in the abstracts. While interpreting the results, it should be noted that the association graphs are generated based on the cumulative evidence counts, which might be biased for a disease or microbe having a higher coverage. In such cases, the individual associations must be carefully assessed using the implemented hypergeometric tests before making any biological inference. The implementation of word cloud for the search output can highlight keywords in the abstracts that get repeatedly mentioned. Although this feature can be used as a tool to understand the mechanism of how the microbes affect various diseases, it is strongly advised to carefully crosscheck with the individual publications. In a future update, we plan to link the results with human genome-wide association studies and other related databases to help users automatically get improved insights. We also plan to augment an additional layer of natural language processing to help users automatically get insights on the nature of interaction in a future update. Additionally, we will introduce a "Contribute" feature to allow users pick a random abstract from an initial preselected set of abstracts and submit their annotation on the observed type of association (both microbe–microbe and microbe–disease). Every annotation will be cross validated by two other independent annotations to improve accuracy. We expect EviMass to serve as a valuable resource for microbiologist as well as other researchers working in the field of human microbiome and diseases.

MATERIAL, METHODS AND IMPLEMENTATION

Data Acquisition and Building the EviMass Backend

Generation of the EviMass backend involved two major steps, namely, information extraction and entity recognition. Articles with abstracts were downloaded directly from PubMed. A combination of keywords including "microbe," "microbiome," "microbial disease," "metagenome," and "bacteria" was used to query abstracts using the PubMed web interface. There were 1,457,991 unique articles retrieved, which were parsed using in-house scripts to retain PubMed IDs, title, publication year, journal name, authors, and abstract text. These abstracts were further processed to extract bacteria names and the reported human diseases. The steps involved in backend processing are described below as well as summarized in **Figure 5**. Processed backend tables along with their description are provided in the **Supplementary data**.

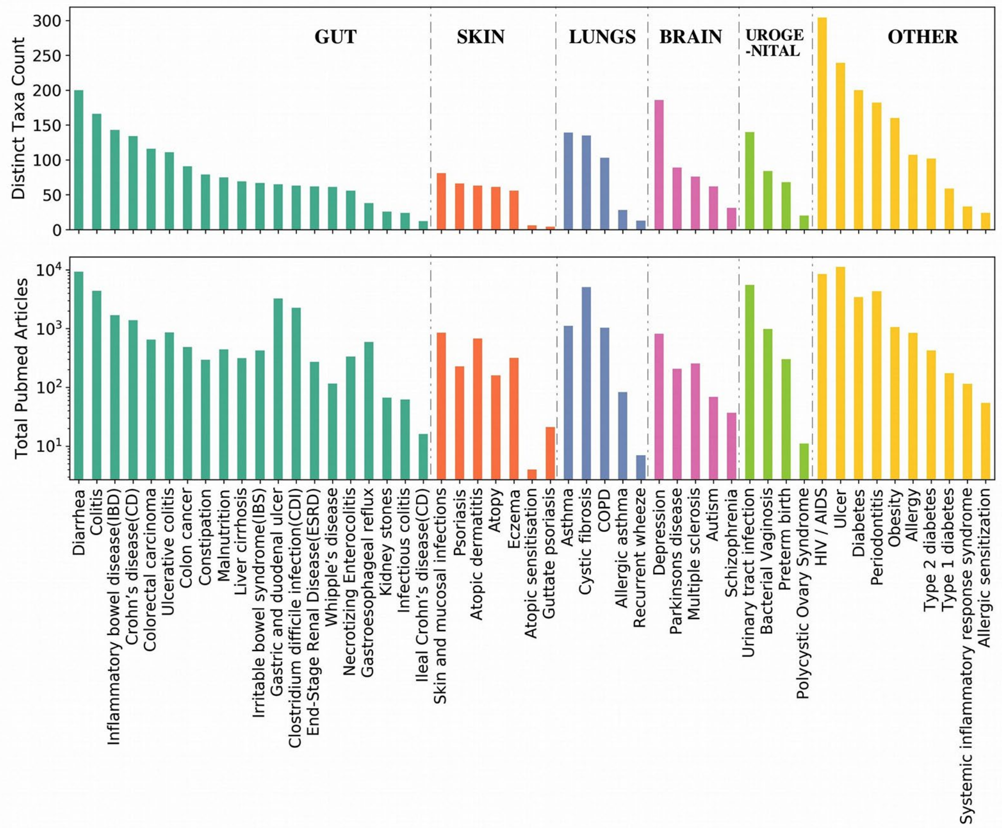


FIGURE 3 | Summary of the associated microbial genera count corresponding to each disease and the number of articles reporting the disease. The diseases are ordered based on the categories as listed in **Table 1**. Each category of disease is sorted based on the number of genera associations.

Bacteria Named Entity Recognition

The abstracts were passed through a named entity recognition (NER) engine implemented in the BacNER tool (Wang et al., 2017a). BacNER is a dedicated bacterial NER tool, which reports bacteria names, strains, and related entities from a given query text. It is based on a trained conditional random field, which processes text and tags bacterial entities in IOB (inside-outside-beginning) format. The title and the abstract for each article were passed to BacNER, and the entities reported in them were extracted. A total of 787,069 articles from our library were returned with at least one bacterial entity recognized. The results from BacNER required further processing in order to be used in our model. For instance, entities like *Escherichia coli* and *E. coli* needs to be clubbed together. Moreover, there were instances where specific species/strains of a bacterium were reported, which needed to be clustered together. The identified species were also kept as a separate map with the PubMed IDs to display them in the EviMass web tool. To resolve these ambiguities, a master list of 2,178 genera was generated using the Ribosomal Database Project (Maidak et al., 1996) and Green Genes (DeSantis et al., 2006) database. As the majority of microbiome 16S rRNA studies utilize one of these databases, it also aligns to our aim of validating the results from microbiome experimental data. Using an approximate string matching method based on Levenshtein distance (Miller et al., 2009), each identified bacterial entity was matched and mapped to the master list. The mapping was then manually verified to modify inconsistent mappings. A total of 637,428 articles were finally selected having a mapped bacterial entity to the biomedical text. A detailed description of the steps involved is summarized in **Figure 5**.

Diseases Named Entity Recognition

In order to create a disease entity dictionary, the HMDAD's most commonly occurring list of diseases (Ma et al., 2017a) was used along with some additions to finalize a set of 51 diseases. The disease set is created in order to effectively cater to the wide variety of researches. For example, "diabetes" is deliberately kept as a different disorder and is not merged with "type 1" or "type 2 diabetes." Another example of a similar case pertains to the disease "colorectal carcinoma" where we added a search query term for both "colorectal carcinoma" and "colorectal cancer" to encompass all the search results. These 51 diseases were further grouped into 6 categories broadly based on their target regions: gut, skin, lungs brain, urogenital, and other (systemic diseases) (**Table 1**). Disease names were recognized from abstracts identified earlier to have an associated bacterial entity using string matching.

The complete information extracted from more than a million scientific articles is stored and indexed for minimum memory consumption and fast access. All the genera as well

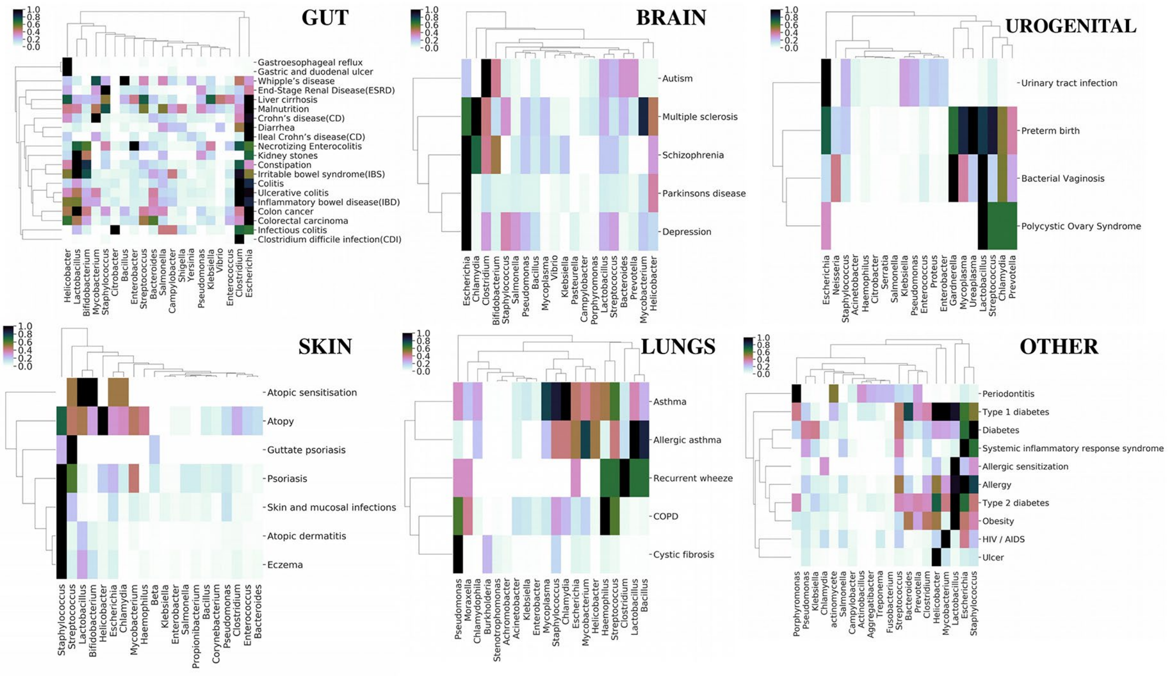


FIGURE 4 | Category-wise (organs affected by various diseases) bidirectionally clustered heat maps based on microbial associations. The top 20 persistent microbes across the six categories (**Table 1**) were chosen and used to generate bidirectionally clustered (UPGMA hierarchical clustering) heat map for each category. Euclidean distance was used as the measure of distance, and the values were normalized by rows (diseases).

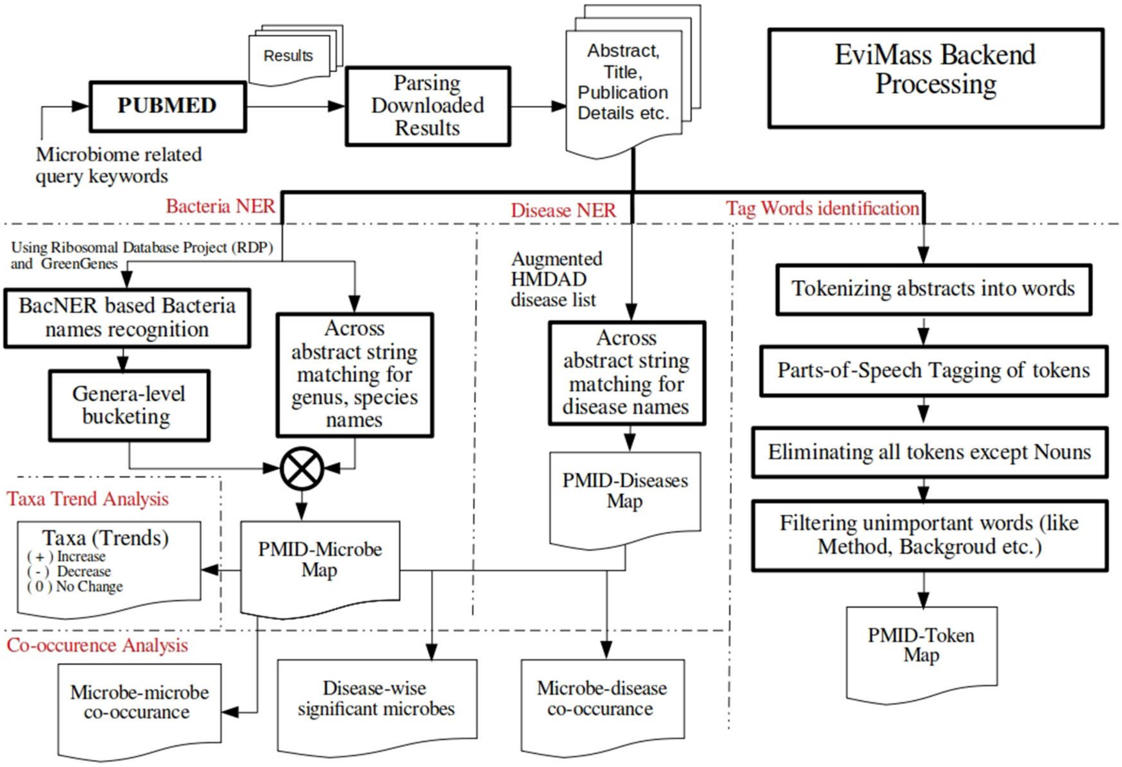

EviMass Backend Processing
Results
PUBMED
Microbiome related query keywords
Parsing Downloaded Results
Abstract, Title, Publication Details etc.
Bacteria NER
Disease NER
Tag Words identification
Using Ribosomal Database Project (RDP) and GreenGenes
BacNER based Bacteria names recognition
Genera-level bucketing
Across abstract string matching for genus, species names
Augmented HMDAD disease list
Across abstract string matching for disease names
PMID-Diseases Map
Tokenizing abstracts into words
Parts-of-Speech Tagging of tokens
Eliminating all tokens except Nouns
Filtering unimportant words (like Method, Backgroud etc.)
PMID-Token Map
Taxa Trend Analysis
Taxa (Trends)
(+) Increase
(-) Decrease
(0) No Change
PMID-Microbe Map
Co-occurence Analysis
Microbe-microbe co-occurance
Disease-wise significant microbes
Microbe-disease co-occurance

FIGURE 5 | Flowchart describing the various steps involved in development of the EviMass backend.

as the diseases reported for articles are stored in tables, where each record corresponds to a PubMed ID. Apart from this, all PubMed IDs that report each genus are also separately identified and stored. Similarly, a mapping of disease and PubMed IDs is also created for easy information retrieval. Abstracts are then processed to identify "increase" and "decrease" of the various microbial names identified to be present in them. These patterns are later displayed in the web application in the PMID result table under "taxa and trends" column with a "+" (increase), "−" (decrease), and 0 (no trend detected) sign beside the identified taxa name in an abstract. For advanced analysis, EviMass holds all the parts-of-speech (POS) tagged noun words corresponding to the articles, which can be used to get a deeper insight. These POS tags can be used to fine tune a search based on a particular term of interest as described in the next section (**Figure 5**). Microbial genera significantly associated with the diseases ($P < 0.05$) were identified using a Fisher exact test (Lo and Marculescu, 2017; Ma et al., 2017a), which is further applied for enrichment analysis in the web tool.

The EviMass Frontend

EviMass web server uses the generated backend to allow easy queries using simplistic searches and graphical outputs. Three workflows are implemented to systematically query for a microbe–microbe or disease–microbe association as described below (additional details in **Supplementary material 2**).

Module 1: Identify Intermicrobial Associations

Using this module, users can select a microbial genus and find all other microbial genera associated with it. The results of the workflow are presented as a network with the central node representing the queried genus and the peripheral nodes representing the associated genera. The sizes of the nodes represent the strengths of the associations and are calculated as the total number of publications where the two genera (corresponding to the central and the peripheral node) are identified to co-occur. EviMass displays the top 100 strongest associated pairs by default but also provide users an option to view all the associations. Along with the network, a dropdown/text box with automatic suggestions for associated microbial genera names is rendered. Clicking on any node or selecting any microbial genera from the dropdown will display all the PMIDs in which the corresponding genera and the queried genera co-occur, along with the main keywords (POS tags) used in the abstract listed as a table. Additionally, a set of hypergeometric tests, namely, Fisher exact test and χ^2 test, are performed (Camilli, 1995; Lo and Marculescu, 2017; Ma et al., 2017a) to statistically assess the significance of the selected association, and the results are presented as a contingency table along with *P* values. Users have the option to search and filter the displayed table for any term/keyword and narrow down the number of abstracts containing the specified word using either the global search or a column-specific search. Also, to ease further analysis, a word cloud of entity names in the abstracts from the PMID resultant output table can be generated for a specific custom query. If a particular gene, protein, or clinical condition gets repeatedly mentioned in the abstract texts for the selected interaction, it will appear as a dominant word. The PMID output table can also be downloaded in a variety of commonly used formats. EviMass also allows users to identify inter microbial associations, which are present in a selected set of diseases using interactive options.

Module 2a: Identify All Microbial Genera Associated With a Disease

This module can be used to find all genera that are reported to be associated with a selected disease. The results of this module can be viewed either as a network (with the central node being the disease and the peripheral nodes being the associated microbial genera) or as a bar chart with the top 30 associated genera sorted by their strength of associations. Microbes identified to be significantly ($P < 0.05$) associated with the selected diseases are highlighted in pink (nodes/bars). In addition, a dropdown/text box with automatic suggestions for associated microbial genera names is provided for convenience. Clicking on any peripheral node (in case of the network view) or bar (in case of bar chart) or selecting any microbial genus from the dropdown displays the PMIDs in which the disease and the corresponding genus co-occur along with the keywords in the abstract in a sortable, searchable, and downloadable table. Similar to Module 1, results for assessing significance of the association are also generated. In addition, a genus node can be interactively queried (using left mouse click) to inquire its other known disease associations as a separate bar plot.

Module 2b: Identify All Diseases Associated With a Microbial Genera

The diseases associated with a particular genus can be evaluated using this module. A genus can be queried to find its associations with the diseases depicted in form of a bar chart, arranged in order of the strength of their associations along with their statistical significance. As in the previous modules, clicking on any bar or selecting any associated disease from the dropdown will load the PMIDs where the corresponding disease and the queried microbe co-occur along with the keywords in a sortable, searchable, and downloadable table. The "word cloud" for the PMID resultant table can be used to understand the mechanism of how the microbe affects the disease (**Supplementary Figure 25**).

Module 3a: View Literature Evidence for a Disease-Specific Microbial Network

Often, biological systems are analyzed as a network/graph, which are mostly obtained using computational techniques on microbiome abundance data. However, such data-driven approaches often lead to spurious connections among noninteracting microbes, due to either measurement or statistical errors. Therefore, a quick and easy method to correlate such associations with literature mined results is likely to help in getting an improved understanding. This module offers users the possibility to upload a microbial association network as an edge list along with the pertinent disease.

The uploaded edge list is depicted as a network with a searchable dropdown containing all the edges. For user convenience, the edge widths are automatically mapped to their association frequencies. Clicking on any edge or selecting any edge from the dropdown shows the PMIDs and keywords where the pair co-occurs along with a list of evidence statistics. The evidence statistics reports the occurrence count of the selected genera independently as well as together in the given disease, any diseases, and globally in the EviMass backend. The utility of this feature has been demonstrated as a case study in the Results section.

Module 3b: View Literature Evidence for Genera Identified to Be Differentially Abundant in a Disease and Perform Enrichment Analysis

Analyzing differentially abundant microbial genera in disease–healthy microbiome studies is often used to identify potential microbial biomarkers. This module enables one to view literature reported evidences for associations of a given set of differentially abundant microbial genera (identified from an experimental study) with a specific disease. The results of the module can be viewed either as a network, with the central node depicting the disease and the peripheral nodes representing the queried microbial genera, or as a bar chart with the queried genera sorted by their strength of associations. In addition, a dropdown/text box with automatic suggestions for associated microbial genera names is rendered. Clicking on any peripheral node (in case of the network view) or bar (in case of bar chart) or selecting any microbial genus from the dropdown displays the PMIDs in which the disease and the corresponding genus co-occur along with the keywords in the abstract in a sortable, searchable, and downloadable table. All the other disease associations of the genus corresponding to the selected node/bar are reported as a separate bar chart. An enrichment analysis of the uploaded set of microbial genera is performed with respect to the selected disease similar to the implementation in Micro-pattern (Ma et al., 2017a). For this implementation, the microbes identified to be significantly associated with the 51 diseases are used as "disease sets" in EviMass.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://web.rniapps.net/netshift/datasets/allergy.zip.

AUTHOR CONTRIBUTIONS

BK conceived the idea. DS extracted and processed/analyzed the data and created the backend. KB designed and developed the web server and implemented the statistical tests. BK, DS, and KB designed the case studies. BK, DS, KB, and SM evaluated the results and drafted the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

KB is an industry-sponsored AcSIR PhD student at Chemical Engineering & Process Development Division, CSIR–National Chemical Laboratory (NCL), Pune 411008 (India), and would like to acknowledge NCL and 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/fgene.2019.00849/full#supplementary-material

REFERENCES

Camilli, G. (1995). The relationship between Fisher's exact test and Pearson's chi-square test: a Bayesian perspective. *Psychometrika* 60, 305–312. doi: 10.1007/BF02301418

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

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G.-Y., and Wang, X.-S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

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

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.* 45, W180–W188. doi: 10.1093/nar/gkx295

Eloe-Fadrosh, E. A., and Rasko, D. A. (2013). The human microbiome: from symbiosis to pathogenesis. *Annu. Rev. Med.* 64, 145–163. doi: 10.1146/annurev-med-010312-133513

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

Fazlollahi, M., Lee, T. D., Andrade, J., Oguntuyo, K., Chun, Y., Grishina, G., et al. (2018). The nasal microbiome in asthma. *J. Allergy Clin. Immunol.* 142, 834–843, e2. doi: 10.1016/j.jaci.2018.02.020

Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., et al. (2014). Conducting a microbiome study. *Cell* 158, 250–262. doi: 10.1016/j.cell.2014.06.037

Hevia, A., Milani, C., López, P., Donado, C. D., Cuervo, A., González, S., et al. (2016). Allergic patients with long-term asthma display low levels of *Bifidobacterium adolescentis*. *PLoS One* 11, e0147809. doi: 10.1371/journal.pone.0147809

Huang, Y.-A., You, Z.-H., Chen, X., Huang, Z.-A., Zhang, S., and Yan, G.-Y. (2017a). Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15, 209. doi: 10.1186/s12967-017-1304-7

Huang, Z.-A., Chen, X., Zhu, Z., Liu, H., Yan, G.-Y., You, Z.-H., et al. (2017b). PBHMDA: path-based human microbe–disease association prediction. *Front. Microbiol.* 8, 233. doi: 10.3389/fmicb.2017.00233

Jiang, S., Hu, Y., Deng, S., Deng, J., Yu, X., Huang, G., et al. (2018). miR-146a regulates inflammatory cytokine production in *Porphyromonas gingivalis* lipopolysaccharide-stimulated B cells by targeting IRAK1 but not TRAF6. *Biochim. Biophys. Acta Mol. Basis Dis.* 1864, 925–933. doi: 10.1016/j.bbadis.2017.12.035

Kajiura, Y., Nishikawa, Y., Lew, J. H., Kido, J.-I., Nagata, T., and Naruishi, K. (2018). β-Carotene suppresses *Porphyromonas gingivalis* lipopolysaccharide-mediated cytokine production in THP-1 monocytes cultured with high glucose condition. *Cell Biol. Int.* 42, 105–111. doi: 10.1002/cbin.10873

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

Kuntal, B. K., Chandrakar, P., Sadhu, S., and Mande, S. S. (2019). 'NetShift': a methodology for understanding 'driver microbes' from healthy and disease microbiome datasets. *ISME J.* 13, 442. doi: 10.1038/s41396-018-0291-x

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

Larsen, J. M. (2017). The immune response to *Prevotella* bacteria in chronic inflammatory disease. *Immunology* 151, 363–374. doi: 10.1111/imm.12760

Liang, D., Leung, R. K.-K., Guan, W., and Au, W. W. (2018). Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathog.* 10, 3. doi: 10.1186/s13099-018-0230-4

Lo, C., and Marculescu, R. (2017). MPLasso: inferring microbial association networks using prior microbial knowledge. *PLoS Comput. Biol.* 13, e1005915. doi: 10.1371/journal.pcbi.1005915

Ma, W., Huang, C., Zhou, Y., Li, J., and Cui, Q. (2017a). MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. *Sci. Rep.* 7, 40200. doi: 10.1038/srep40200

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017b). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., and Woese, C. R. (1996). The Ribosomal Database Project (RDP). *Nucleic Acids Res.* 24, 82–85. doi: 10.1093/nar/24.1.82

Miller, F. P., Vandome, A. F., and McBrewster, J. (2009). *Levenshtein distance: information theory, computer science, string (computer science), string metric, Damerau–Levenshtein distance, spell checker, hamming distance*. Indianapolis, Indiana, United States: Alpha Press.

Peng, L.-H., Yin, J., Zhou, L., Liu, M.-X., and Zhao, Y. (2018). Human microbe–disease association prediction based on adaptive boosting. *Front. Microbiol.* 9, 2440. doi: 10.3389/fmicb.2018.02440

Qu, J., Zhao, Y., and Yin, J. (2019). Identification and analysis of human microbe–disease associations by matrix decomposition and label propagation. *Front. Microbiol.* 10, 291. doi: 10.3389/fmicb.2019.00291

Wang, X., Jiang, X., Liu, M., He, T., and Hu, X. (2017a). "Bacterial named entity recognition based on dictionary and conditional random field," in *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)* (Kansas City, MO, USA: IEEE), 439–444. doi: 10.1109/BIBM.2017.8217688

Wang, F., Huang, Z.-A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017b). LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* 7, 7601. doi: 10.1038/s41598-017-08127-2

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe–disease associations inference based on network consistency projection. *Sci. Rep.* 8, 8034. doi: 10.1038/s41598-018-26448-8

Zhou, Y., Zhang, H., Zhang, G., He, Y., Zhang, P., Sun, Z., et al. (2018). Calcitonin gene-related peptide reduces *Porphyromonas gingivalis* LPS-induced TNF-α release and apoptosis in osteoblasts. *Mol. Med. Rep.* 17, 3246–3254. doi: 10.3892/mmr.2017.8205

Conflict of Interest Statement: All the authors are employed by company Tata Consultancy Services Limited and declare no conflict of interest.

CORRECTION
published: 02 November 2020
doi: 10.3389/fgene.2020.614051

Corrigendum: "EviMass": A Literature Evidence-Based Miner for Human Microbial Associations

Divyanshu Srivastava [1†], *Krishanu D. Baksi* [1,2†], *Bhusan K. Kuntal* [1,3,4*] and *Sharmila S. Mande* [1*]

[1] Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune, India, [2] School of Information Technology, Indian Institute of Technology Delhi, Delhi, India, [3] Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune, India, [4] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

Keywords: microbiome, literature mining, Human Disease, Web server, microbial association

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

****Correspondence:***
Bhusan K. Kuntal
kuntal.bhusan@tcs.com
Sharmila S. Mande
sharmila.mande@tcs.com

†These authors have contributed equally to this work

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

Received: *05 October 2020*
Accepted: *06 October 2020*
Published: *02 November 2020*

Citation:
Srivastava D, Baksi KD, Kuntal BK and Mande SS (2020) Corrigendum: "EviMass": A Literature Evidence-Based Miner for Human Microbial Associations. Front. Genet. 11:614051. doi: 10.3389/fgene.2020.614051

A Corrigendum on

"EviMass": A Literature Evidence-Based Miner for Human Microbial Associations
by Srivastava, D., Baksi, K. D., Kuntal, B. K., and Mande, S. S. (2019). Front. Genet. 10:849. doi: 10.3389/fgene.2019.00849

In the published article, there was an error in affiliation **4**. Instead of "**Academy of Scientific and Innovative Research (AcSIR), CSIR-National Chemical Laboratory Campus, Pune, India"** it should be "**Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India"**.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

REVIEW
published: 25 September 2019
doi: 10.3389/fgene.2019.00904

Advances and Challenges in Metatranscriptomic Analysis

Migun Shakya, Chien-Chi Lo and Patrick S. G. Chain**

Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States

Sequencing-based analyses of microbiomes have traditionally focused on addressing the question of community membership and profiling taxonomic abundance through amplicon sequencing of 16 rRNA genes. More recently, shotgun metagenomics, which involves the random sequencing of all genomic content of a microbiome, has dominated this arena due to advancements in sequencing technology throughput and capability to profile genes as well as microbiome membership. While these methods have revealed a great number of insights into a wide variety of microbiomes, both of these approaches only describe the presence of organisms or genes, and not whether they are active members of the microbiome. To obtain deeper insights into how a microbial community responds over time to their changing environmental conditions, microbiome scientists are beginning to employ large-scale metatranscriptomics approaches. Here, we present a comprehensive review on computational metatranscriptomics approaches to study microbial community transcriptomes. We review the major advancements in this burgeoning field, compare strengths and weaknesses to other microbiome analysis methods, list available tools and workflows, and describe use cases and limitations of this method. We envision that this field will continue to grow exponentially, as will the scope of projects (e.g. longitudinal studies of community transcriptional responses to perturbations over time) and the resulting data. This review will provide a list of options for computational analysis of these data and will highlight areas in need of development.

Keywords: RNASeq, microbiome, workflows, gene expression, omics

OPEN ACCESS

Edited by:
Bas E. Dutilh,
Utrecht University, Netherlands

Reviewed by:
Alejandro Sanchez-Flores,
National Autonomous
University of Mexico, Mexico
Guilherme Corrêa De Oliveira,
Vale Technological Institute (ITV),
Brazil

****Correspondence:***
Migun Shakya
migun@lanl.gov
Patrick S. G. Chain
pchain@lanl.gov

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

Received: *15 May 2019*
Accepted: *26 August 2019*
Published: *25 September 2019*

Citation:
Shakya M, Lo CC and Chain PSG (2019) Advances and Challenges in Metatranscriptomic Analysis. Front. Genet. 10:904. doi: 10.3389/fgene.2019.00904

INTRODUCTION

The past few decades have seen significant advancements in sequencing technologies that have transformed how we conduct biological experiments, particularly when it comes to the study of complex microbiomes. However, most of the high throughput sequencing has focused on DNA sequencing of entire communities using either targeted approaches like PCR-amplicon sequencing of 16S rRNA genes (or other marker genes) or shotgun sequencing of all available DNA from the sample (metagenomics).

These methods have contributed to many discoveries in the past decade, helping to better characterize microbiomes from environments ranging from the human gut (Qin et al., 2010) to soil (Rondon et al., 2000) to oceans (Venter et al., 2004). Although 16S studies only directly characterize the taxonomic profile of a microbiome, it is a cost-effective option to exhaustively capture biodiversity (measuring the maximal dynamic range of relative abundance) of many samples using minimal sequencing. However, more and more studies are now using shotgun metagenomics as the advancements in sequencing technologies allow the comprehensive capture of most microbiome members while at the same time elucidating potential genes and functional pathways. One of the main limitations of shotgun metagenomics is that it does not distinguish the active from inactive members of a microbiome, and thus cannot help discriminate those that are contributing to observed ecosystem behavior from those that are merely present, presumably awaiting more favorable conditions.

Using RNA sequencing (RNASeq) to record expressed transcripts within a microbiome at a given point in time under a set of environmental conditions provides a closer look at active members. Recent advancements in mass spectrometry methods applied towards proteomics is also able to provide insight into actively expressed proteins, but is best paired with known reference genomes or a reference metagenome from which expected peptide masses can be matched. With RNASeq, relatively lowly expressed genes including the entire metatranscriptome that include non-coding RNAs can be detected, annotated, and mapped to metabolic pathways.

Biologists have long measured RNAs using targeted approaches like qPCR to quantify expression of known genes of interest. Before the advent of high throughput sequencing, microarray technologies were also widely used to measure the expression levels of known transcripts from organisms or even communities (Parro et al., 2007). With the application of next-generation sequencing (NGS) technologies to RNA, it is now possible to not only measure known transcript targets but also discover previously unknown transcripts and transcript variants directly from the sequence data.

In the short time since it was first introduced in the early 2000s, the number of metatranscriptomics projects, or the sequencing of RNAs from microbial communities has increased significantly (**Figure 1**). In terms of applications, the technique has been used to characterize active microbes in a community (Bashiardes et al., 2016), discover novel microbial interactions (Bikel et al., 2015), detect regulatory antisense RNA (Bao et al., 2015), and track expression of genes and determine the relationship between viruses and their host (Moniruzzaman et al., 2017). This revolutionary method is not a complete panacea however, and comes with its own set of drawbacks. As with most transcriptomic methods, experimental design is critical, sample collection is destructive and sufficient material for sequencing (or coupled experiments) is required. In addition, metatranscriptomics is not always able to capture the entire metatranscriptome due in part to the complexity (high diversity and relative ratios of members) of some microbial communities, the large dynamic range of transcript expression, the short half-life of RNA, and a number of technology-specific limitations.

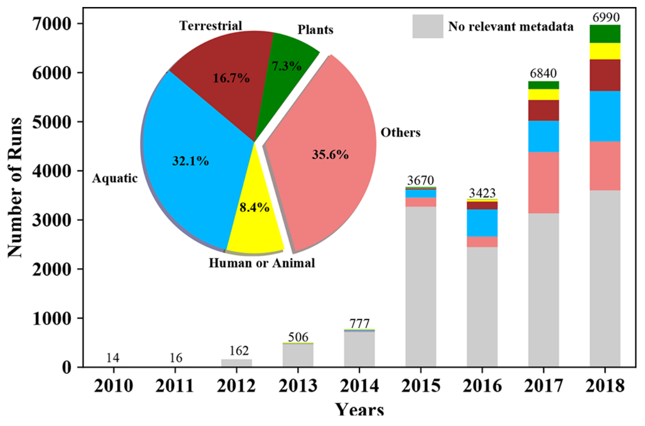


FIGURE 1 | Growth of metatranscriptomics projects in public repositories, together with associated metadata, over time. Bars plots represent number of metatranscriptomic datasets (i.e. "runs") deposited in the NCBI Sequence Read Archive (SRA) on a per annual basis. The pie chart and the stacked bars are colored based on the source/environment (isolation_source) the sample has been isolated from. The lowest bar in grey represents the number of samples in SRA without this pertinent metadata.

In this review, we report the state of metatranscriptomics by discussing several microbiome studies from different ecosystems. We will discuss both novel findings made possible by this methodology as well as some of the shortcomings. We also list several of the available tools and workflows that have been adopted for or have been specifically designed to analyze metatranscriptomic datasets.

APPLICATION OF METATRANSCRIPTOMICS ACROSS ECOSYSTEMS

Metatranscriptomics has been applied to a number of different types of samples, from the study of human (and other animal) microbiomes, to those found in or on plants, within soils, and in aquatic environments. In this section we provide some examples of the impact metatranscriptomics has had in different fields of study.

Aquatic Environments

One of the first metatranscriptomic studies was conducted on freshwater bacterioplankton communities (Poretsky et al., 2005), which described a total of 400 environmental transcripts from two sites. At the time, the scale of the study was dictated by the available sequencing technologies that limited the sensitivity of the method to only a few hundred genes. With the advent in the high throughput sequencing technologies, other studies on marine systems produced hundreds of thousands of reads per sample (Frias-Lopez et al., 2008; Gilbert et al., 2008) and made it possible to use metatranscriptomics to characterize the dynamics of cyanobacterial blooms in the Baltic sea (Berg et al., 2018), the detection of small RNAs in the open ocean (Shi et al., 2009), and resolve viral-host relationships of marine eukaryotes (Moniruzzaman et al., 2017).

Terrestrial Environments

Soils are some of the most diverse ecosystems in the world. They typically harbor incredible numbers and a broad diversity of eukaryotes, archaea, bacteria, and viruses. These complex microbiomes are frequently characterized using metagenomic sequencing, but only a few of studies have performed metatranscriptomics to decipher active microbes from more sedentary soil residents. For example, in a recent study to identify functionally active organisms in soil microbial communities, metatranscriptomes revealed that *Verrucomicrobia,* which are regularly found in high abundance in soils, were not as highly active as their abundance would otherwise suggest (White et al., 2016). Upon further analyses, authors showed that the high abundance of *Verrucomicrobia* at DNA level was partly due to presence of metabolically inactive organisms. Since it is possible to sequester eukaryotic mRNA during sample preparation (e.g. using polyA tail hybridization), metatranscriptomics allows the targeting of just eukaryotic mRNA. Using this approach, a survey of forest soils helped characterize the taxonomic diversity and also discovered genes that code for novel eukaryotic Carbohydrate-Active enzymes (Damon et al., 2012). Likewise, the large diversity of active protists in mineral and organic soils were identified using the approach (Geisen et al., 2015). Going forward, metatranscriptomics will be pivotal in characterizing diversity of active soil organisms and functions.

Human Microbiomes

In the past decade, our understanding of the human microbiome has rapidly expanded thanks to sequencing technologies that made possible the description of human gut microbial diversity across large human cohorts (Arumugam et al., 2011; Human Microbiome Project, 2012). Although past studies have primarily focused on describing the taxonomic composition of microbial communities and their functional potential, many studies are now also using metatranscriptomics to better understand the interactions among microbes and their host (Pérez-Losada et al., 2015), to identify active pathways of importance (Franzosa et al., 2014), and how expressed functions may impact disease progression (Nowicki et al., 2018) and severity (Schirmer et al., 2018). A longitudinal study of Inflammatory Bowel Disease (IBD) showed that two organisms *Alistipes putredinis* and *Bacteroides vulgatus* were the sole contributors to the expression of methylerythritol phosphate pathway at different time points. Interestingly, expression by specific organisms correlated with disease severity as *A. putredinis* showed negative and *B. vulgatus* showed a positive correlation (Schirmer et al., 2018). With further advancements in sequencing technologies, laboratory protocols and chemistry, and tailored bioinformatic analysis methods, metatranscriptomics promises to become an integral tool to investigate microbiomes in humans.

Additional Animal-Microbe Interactions

Metatranscriptomic approaches have also been adapted to better understand the microbiomes of other animals, such as cattles (Mann et al., 2018; Sollinger et al., 2018; Li et al., 2019), squirrels (Hatton et al., 2017), and birds (Marcelino et al., 2019). Many studies in cattle microbiomes are focused on understanding the rumen microbiota to mitigate the release of potent greenhouse gas methane from livestock and increase feed efficiency. Through the use of metatranscriptomics, studies have linked microbes in the rumen to pertinent activities such as methane emission and the degradation of complex plant polysaccharides. For example, Sollinger et al. (2018) found *Prevotella* of *Bacteroidetes* and multiple members of *Firmicutes* were actively involved in the degradation of complex saccharides.

Plant-Microbe Interactions

Metatranscriptomics has been applied to many plant-microbe interactions studies as it is able to characterize members of a microbiome that are responsible for specific functions and elucidate genes that drive the relationship of the microbiome with its host. Metatranscriptomic sequencing of all community members from roots of the willow plant *Salix purpurea* cv. Fish" Creek" grown in soil contaminated with petroleum hydrocarbons revealed that the bacterial symbiont *Enterobacteriaceae* was responsible for the degradation of hydrocarbons from among a wide range of active microbes (Gonzalez et al., 2018). The approach is also well suited to detect changes in the microbial community that would have been missed by traditional PCR methods as shown in a study where an increase in diversity of non-fungal eukaryotes was detected in *sad1* mutant of oat plants when compared to its wild type (Turner et al., 2013). The methodology also helped to identify genes that are responsible for the mutualistic relationship of the Seagrass plant with its microbiome members (Crump et al., 2018) and to describe the active microbial communities and pathways in mature ripe fruits (Saminathan et al., 2018). Another example of an attempt to understand mechanisms behind the suppressive and non-suppressive *Rhizoctonia solani* fungal infection in wheat plants revealed a set of genes associated with suppression and non-suppression phenotypes, providing molecular targets for improved agricultural productivity (Hayden et al., 2018).

BIOINFORMATIC ANALYSIS OF METATRANSCRIPTOMIC SEQUENCING DATA

Because of microbiome complexity, high throughput sequencing in the form of short read data usually generated from Illumina sequencing technology has been most often applied for metatranscriptome studies, particularly when multiple samples and deep coverage are required, such as in differential gene expression studies. Since most information about samples are unknown *a priori*, such as its microbial composition, relative abundance of community membership, genome sizes, and relative expression within and among genomes, it is not trivial to find right experimental parameters such as depth of sequencing for metatranscriptomics. While long read sequencing can produce full or near full-length mRNAs which can help discriminate among different isoforms (Pollard et al., 2018), and provide longer stretches of sequence for similarity searches, the various long read technologies currently only play a supporting role and are not actively being used alone for metatranscriptome studies. Here, we focus on available tools and workflows for metatranscriptome data processing and analysis, which focus on short read data.

Preprocessing

Similar to other NGS datasets, one of the first steps in processing RNASeq data is to do Quality Control (QC) and remove or trim spurious/erroneous reads to minimize errors. One of the many dozens of available QC tools, such as FastQC (Andrews, 2010), FaQCs (Lo and Chain, 2014), fastp (Chen et al., 2018), and Trimmomatic (Bolger et al., 2014), can be used for short reads derived from Illumina sequencers.

One of the important steps that should be taken into consideration is physical removal or depletion of the highly abundant ribosomal RNA (rRNA) transcripts from the samples, as they often constitute upward of 90% of all data if not removed and do not contribute towards most downstream analyses, such as finding differentially expressed genes or pathway characterization. These rRNAs are often removed using molecular approaches prior to sequencing but their dominance in samples results in some amount of rRNA still being sequenced. Post sequencing, rRNAs can be identified for removal from downstream analyses using tools like SortMeRNA (Kopylova et al., 2012) and barrnap (Seemann, 2014).

There are also cases where one would want to remove a target organism from analysis, such as human reads from human microbiome samples. These reads can be removed using traditional read mapping methods that tags and removes reads that map to human genome (Li et al., 2017), or using faster alignment free methods such as Best Match Tagger (BMTagger) (Rotmistrovsky and Agarwala, 2011) that search for human-specific *k*-mers in reads.

De Novo Assembly

Preprocessed, high-quality reads can now be assembled into putative transcripts using *de novo* assemblers. Given that most microbiomes are not adequately characterized with reference genomes, *de novo* assemblers provide a reference scaffold representing longer, expressed genome segments that can provide a reference set of genes. This provides users the ability to find homologs in a more straightforward fashion, establish taxonomic origin, and serve as a reference for mapping against for expression analysis. Metagenomic assemblers such as MEGAHIT (Li et al., 2015), IDBA-UD (Peng et al., 2012) and metaSPAdes (Nurk et al., 2017) have been designed to tackle complex metagenomes that may share some sequence similarity in highly conserved regions but may vary greatly in terms of relative abundance within the microbiome, and may also harbor population (strain-level) variation. However, the effectiveness of these assemblers in reconstructing transcripts that have their own peculiarities such as introns/exons, different isoforms, and shorter non-coding RNAs (ncRNA), have been seldomly tested, so, it is with caution that one should use metagenomic assemblers on metatranscriptome datasets.

Assemblers such as Trans-ABySS (Robertson et al., 2010), Trinity (Grabherr et al., 2011), BinPacker (Liu et al., 2016), Oases (Schulz et al., 2012), SOAPdenovo-Trans (Xie et al., 2014), IDBA-Tran (Peng et al., 2013), and rnaSPAdes (Bushmanova et al., 2019) attempt to account for the issues in transcriptome sequencing, but were originally designed to assemble transcripts from a single organism. Despite their design towards transcriptomic and not metatranscriptomic datasets, comparisons among some assemblers showed that in general, the tested assemblers Oases, Trinity, Metavelvet, all improved the number of annotated genes from the resulting contigs, with the Trinity assembler outperforming the others (Celaj et al., 2014).

IDBA-MT (Leung et al., 2013), IDBA-MTP (Leung et al., 2014), and Transcript Assembly Graph (TAG) (Ye and Tang, 2016) are *de novo* assemblers that are designed specifically for metatranscriptomes and take into account the unique features of both transcripts and the complex nature of microbial communities. IDBA-MT is built upon IDBA-UD and uses multiple *k* values in a de Bruijn graph while accounting for features associated with mRNAs like uneven sequencing depth and common repeat patterns across different mRNAs, thereby lowering the rate of mis assemblies. Likewise, IDBA-MTP was derived from IDBA-MT to be able to assemble lowly expressed mRNAs. It uses the information of known protein sequences to guide the assembly by starting with smaller *k*-values to construct mRNA sequences which are then included based on their similarity with a known set of proteins. TAG is a comparatively recent assembler that also uses a de Bruijn graph, but to assemble the corresponding metagenome, which is then used as a reference to map the transcriptome reads and reconstruct mRNA sequences by traversing the metagenome assembly graph with mapped transcriptome reads. Since it assumes genes are contiguous (without splicing), this particular tool is ineffective to use in microbiomes that also contain eukaryotes. Furthermore, there is an implicit assumption that the metagenome represents sufficient breadth of the community that all expressed genes can be mapped to the metagenome.

The current state of *de novo* assembly for metatranscriptomic datasets is still in its very early stages. Only a handful of tools have been specifically developed for metatranscriptomics, but their efficacy on diverse datasets has not been tested and their hardware, or memory requirements across an array of community complexities and data volume, have also not been rigorously established.

Transcript Taxonomy

Similar to the taxonomic profiling that is frequently performed with shotgun metagenomic data, one can use the same suite of tools to perform read- or contig-based taxonomic assignments in order to understand what organisms are actively expressing RNA. A separate and distinct method is to focus solely on rRNAs to assess active members of a community, though as mentioned above, these are frequently removed (both in the wet-lab protocols as well as in preprocessing of the raw data).

Read-based taxonomy classification tools such as Kraken (Wood and Salzberg, 2014), GOTTCHA (Freitas et al., 2015), MetaPhlan2 (Truong et al., 2015), etc. can be used for metatranscriptomes (Neves et al., 2017). Because these tools work on short reads and are based on nucleotide matches, their effective use is limited to microbiomes whose members have close neighbors in existing sequence databases. Reads that have been assembled into longer contigs and possibly full-length transcripts can be used by a number of tools, such as centrifuge (Kim et al., 2016a) and Kraken 2 (Wood and Salzberg, 2014), to potentially identify a greater proportion of the sequenced community members.

Taxonomic assignments using reads or predicted coding regions have a large number of limitations, including the algorithms necessary to process large volumes of data or accommodate short sequences, and the paucity of references in the reference databases. Compounding such issues, is the fact that most bioinformatics tools only utilize a subset of available genomes or focus on certain organisms. For example, many tools do not have eukaryotes as part of their databases. There have been some recent efforts with new tools and improvements in existing tools, to include eukaryotic genomes within their databases, such as MetaPhlan2 (Truong et al., 2015) and kaiju (Menzel et al., 2016), but their efficacy in classifying eukaryotes is unknown. Furthermore, it is often difficult to discern low abundance hits from false positive hits, which is an innate problem with microbiome studies. Our general lack of knowledge on overall microbial diversity and in any biological system under study can also limit the utility of taxonomy classification tools.

Functional Annotation

One of the main goals of metatranscriptomics is to assess the functional activity of a microbiome. Since the expressed transcripts represent a proxy to the actual phenotype, characterizing the function of transcripts is a fundamental task for metatranscriptomics. Functional annotation can be conducted using either reads or assembled contigs. Read based functional profilers such as MetaCLADE (Ugarte et al., 2018), HMM-GRASPx (Zhong et al., 2016), and UProC (Meinicke, 2015) use tool-specific databases and require predicted open reading frames as input, from other tools like FragGeneScan (Rho et al., 2010). MetaCLADE is one of the latest tools and uses a database that consists of 2 million probabilistic models derived from 15,000 Pfam domains, thus hundreds of models representing any single domain, to encompass the diversity of each domain across the tree of life. A search against this database results in large numbers of hits per read which are then filtered based on redundancy, probability and bit-scores (Ugarte et al., 2018).

Alternatively, annotation of genes can be performed from assembled contigs. Annotation of assembled transcripts proceeds similar to the annotation of genomes and metagenomes. Gene finding using programs like Prodigal (Hyatt et al., 2010) and FragGeneScan (Rho et al., 2010) is followed by functional assignment based on similarity searches using tools such as DIAMOND (Buchfink et al., 2015) to search against functional databases like KEGG (Kanehisa and Goto, 2000), NCBI RefSeq (O'leary et al., 2016), UniProt (Uniprot, 2019) etc. Other tools, pipelines and platforms encompass an array of bioinformatics utilities (including gene finding and annotation), such as Prokka (Seemann, 2014), EDGE Bioinformatics (Li et al., 2017), and MG-RAST (Wilke et al., 2016), which combine a number of similarity searches against various databases, or can even couple assembly, gene calling, and annotation *via* similarity searches. Once annotations are performed, enzymatic functions may also be mapped to known metabolic pathways, using tools like MinPath (Ye and Doak, 2009) or iPath (Yamada et al., 2011).

Differential Expression Analyses

Beyond the simple description of who are the active members and what genes are being expressed at a single time point, are studies of differential gene expression, where metatranscriptomics can be used to compare differing conditions and environmental parameters and their effect on community function or to observe community dynamics over time. There are many tools originally developed for use with single genomes that can be leveraged for metatranscriptomic differential gene expression studies. These tools require as input abundance data per gene (or transcript) and per sample (representing expression under a specific condition or a specific time point). Abundance can be attained in a number of ways, but typically involves some form of read alignment/mapping to a reference genome, a reference assembly or a reference gene set. EdgeR (Robinson et al., 2010), DeSeq2 (Love et al., 2014), and limma (Ritchie et al., 2015) are R packages that are frequently used, together with the abundance information, to identify genes that are statistically significantly differentially expressed among a number of samples (i.e., conditions/timepoints). Likewise, tools such as Generally Applicable Gene-Set/Pathway Analysis (GAGE) can be used to identify pathways enriched in one condition over another (Luo et al., 2009). Since, replicating metatranscriptomics samples are not trivial compared to transcriptomic studies with isolate organisms, non-parametric methods as the implementation in NOISeq (Tarazona et al., 2015) should also be considered.

There are peculiarities in metatranscriptomic analyses that makes differential expression analyses rather challenging, mainly as a result of sequencing a large diversity of transcripts (from a wide array of organisms). Problems such as shared genes among closely related organisms and variation in the taxonomic composition of transcripts can result in incorrect assessment of gene expression profiles. A normalization method has been recently proposed that can minimize the influence of taxonomic diversity in the sample by normalizing count data based on taxonomic composition across different samples, but this method is also biased by representation in taxonomic databases (Klingenberg and Meinicke, 2017).

AVAILABLE WORKFLOWS FOR METATRANSCRIPTOMIC ANALYSIS

As alluded to above, the analysis of a metatranscriptome dataset is laden with choices of bioinformatic steps with many options for tools for any given step. Which steps and tools should be selected are often dictated by the goals of the experiment, the details of which can grow in complexity based on the specifics of the study. However, there do exist bioinformatic workflows that aim to streamline some of this complexity by connecting multiple individual tools into a workflow that can take raw sequencing reads, and process them providing data files that represent the outputs results characterizing taxonomic identities, functional genes, and/or differentially expressed transcripts. Here we summarize eight of the available workflows, namely MetaTrans (Martinez et al., 2016), COMAN (Ni et al., 2016), FMAP (Kim et al., 2016b), SAMSA2 (Westreich et al., 2018), HUMAnN2 (Franzosa et al., 2018), SqueezeMeta (Tamames and Puente-Sánchez, 2018), IMP (Narayanasamy et al., 2016), and MOSCA (Sequeira et al., 2019). We compare the types of analyses these workflows are capable of performing, which dictates what types of biological questions may be addressed using them. Details of these eight workflows, their capabilities (e.g. QC, assembly, differential gene expression analysis), and the specific bioinformatics tools that they use, can be found as a summary in **Table 1** and in detail in **Supplementary Table 1**.

Almost all eight workflows include a form of preprocessing or quality control of raw data, with the exception of HUMAnN2. All the other workflows, aside from FMAP, include as part of this process the removal of reads matching rRNA prior to other analyses. However, FMAP and IMP allows for the targeted removal of host sequences. After the preprocessing step, all workflows essentially take one of two different approaches, either directly using the reads to perform further analyses, or first performing an assembly and annotation, and then using the annotated genes from that assembly for further analyses (**Supplementary Table 1**). MetaTrans, COMAN, FMAP, SAMSA2, HUMAnN2 all use a read-based approach, while SqueezeMeta, IMP, and MOSCA assemble reads into transcripts before further analyses are performed.

Among all read based workflows, MetaTrans is the only one that first detects putative genes prior to further analyses. All other workflows directly use the filtered reads for similarity searches against taxonomic and functional databases. MetaTrans is also unique in that it utilizes the rRNA sequences that were sequestered in previous step for taxonomic profile analysis. FMAP does not perform taxonomy profiling; and all other workflows use the processed reads to query against a reference database. For these workflows, there are however major differences in how each workflow determines the taxonomy profile. COMAN and SAMSA2 perform their read-based searches in a protein space using DIAMOND, albeit using different reference databases, while HUMANn2 uses MetaPhlan2, which performs searches in nucleotide space. While amino acid based searches allow the detection of organisms distantly related to those in the reference database, they are prone to false discovery. In contrast, nucleotide searches are more specific but are unable to identify sequences insufficiently conserved.

For functional characterization using reads, all five read-based workflows use different algorithms to search for functional similarity using different databases. Only MetaTrans performs these searches in nucleotide space, while all other workflows use read-based predicted peptides as queries. All of the available workflows, aside from SAMSA2, also map predicted proteins onto known pathway maps. Analyses of functional profiles of metatranscriptomes using one of these workflows should be carefully interpreted based on how functions are assigned. For example, functional assignments using searches in nucleotide space, especially for proteins coding genes are likely to be less effective if no near neighbors exist in the reference databases.

In comparison to read-based analyses, assembly-based workflows harbor an extra analytical step, where all the reads are first assembled into larger contigs, which can help reduce the size of the data that needs to be processed for further analyses and increases the contiguous length of the expressed transcripts allowing for more accurate searches. All three of the assembly-based workflows provide multiple assembly tools to choose from, however, IMP has an input requirement, a metagenome dataset that corresponds to the same (or similar) sample as the metatranscriptome. The metagenomic data is used together with the metatranscriptome data for co-assembly. The value of combining metagenome and metatranscriptome dataset is that the assembly becomes more representative of the actual community. IMP uses a corresponding metagenome dataset to create better references through iterative assembly of metagenomes and metatranscriptomes. Both SqueezeMeta

TABLE 1 | A list of metatranscriptomics pipelines and their capabilities.

		Read based					Assembly based		
		MetaTrans	**COMAN**	**FMAP**	**SAMSA2**	**HUMAnN2**	**SqueezeMeta**	**IMP**	**MOSCA**
Preprocessing	QC	✓	✓	✓	✓	×	✓	✓	✓
	Removes host reads	×	×	✓	×	×	×	✓	×
	Removes rRNA	✓	✓	×	✓	×	✓	✓	✓
de novo Assembly		×	×	×	×	×	✓	✓	✓
Binning		×	×	×	×	×	✓	✓	×
Taxonomic Profiling	Reads	✓	✓	×	✓	✓	×	×	×
	Contigs	×	×	×	×	×	✓	✓	✓
Functional Annotation	Reads	✓	✓	✓	✓	✓	×	×	×
	Contigs	×	×	×	×	×	✓	✓	✓
Pathway Analysis		✓	✓	✓	×	✓	✓	✓	×
Requires Metagenomes		×	×	×	×	×	×	✓	×
Summary Report		×	×	×	×	×	×	✓	×
Web Interface		×	✓	×	×	×	×	×	×
Multiple Sample Comparisons		✓	✓	✓	✓	✓	✓	×	✓
Differential Expression		✓	✓	✓	✓	×	×	×	✓
Docker		×	×	×	×	✓	×	✓	✓
Conda		×	×	×	×	✓	×	✓	×
Long Read Support		×	×	×	×	×	✓	×	×
Public Code Repository		✓	×	✓	✓	✓	✓	✓	✓

and IMP can, in addition, perform post-assembly contig-binning to help group together contigs (i.e. transcripts) into bins representing the same taxon (i.e. genes expressed from the same genome/species). In all three assembly-based workflows, the final contigs are processed to find genes, to perform taxonomy classification with those genes, and to assign them a function.

While all workflows use the identified genes as a query against a reference protein database for taxonomic classification purposes, each workflow uses a different strategy. The reference databases used are different (e.g. Uniprot vs NR), and each workflow assigns taxonomy using different algorithms and scoring thresholds (i.e. last common ancestor vs best hit). The SqueezeMeta workflow also uses the rRNA reads that were extracted during the preprocessing step to provide an additional community profile. One major drawback that is common among several workflows is the implementation of an unorthodox approach of assigning taxonomy by searching against databases that are designed for functional characterization.

For functional annotation, the IMP workflow simply uses the output of the Prokka pipeline that was used for gene identification and annotation. The MOSCA workflow uses the output of the taxonomic search against Uniprot and assigns functional annotation based on best hit, while SqueezeMeta performs additional Hidden Markov Model searches against several protein family databases. The SqueezeMeta and IMP workflows also provide pathway analysis based on the annotated functions.

Because one of the primary goals of metatranscriptome analyses is to obtain a relative quantification of gene expression, all read-based and assembly-based workflows provide some form of per gene coverage and/or abundance metric (e.g. raw count per gene, or number of reads per kb per million reads sequenced). These abundance values can be used with additional tools to compare relative gene expression between growth conditions or during time-course experiments, the purpose of which is often to help understand what genes and pathways may be important for a particular phenotype under study. For these types of studies, replicate experiments are often required to obtain statistically significant results, thus the relative gene abundance comparisons is often a comparison among many different samples that include several biological replicates. MetaTrans, FMAP, COMAN, and MOSCA innately provide such a comparative capability within their workflows, can process several datasets and generate a list of genes that are found to be statistically significantly differentially expressed between different conditions (or time points). SAMSA2 also allows differential gene expression analysis but requires individual sample processing followed by the use of an additional command line utility provided as part of the package.

All workflows, with the exception of COMAN, provide a code repository and is invoked using Command Line Interface. COMAN provides a web server interface. The availability of multiple workflows enables users to choose the one that is the most appropriate for analyzing their metatranscriptome. While users should ideally select workflows based on capability/functionality and quality of the algorithms/tools used, additional considerations may include the computational resource requirements, which vary among workflows, and the frequency of maintenance or active development of the source code, which can undergo frequent modifications as new advances, tools, or methods continue to be developed. Both **Table 1** and **Supplementary Table 1** are compilations of these available workflows and can be used as a potential guide to choose a workflow based on factors that are important to address any researcher's question(s). For example, if differential expression analysis is the goal of a study, the list of workflows to choose from is limited to five.

METATRANSCRIPTOMICS—A FUTURE FULL OF PROMISES AND CHALLENGES

As alluded to above, it is clear that the next generation sequencing revolution that has taken place in the study of genomes and metagenomes has been successfully adapted to the study of gene expression with ”RNAseq,” and further, to the study of complex biological system dynamics with metatranscriptomics. This new field has seen a rapid increase in the number of metatranscriptomic projects, most of which represent differential gene expression studies whose goals include obtaining insight into the active members, genes, and pathways within a microbiome. That goal, however, is plagued by the lack of adequate reference genomes, which can result in a suboptimal fraction of reads from any dataset from being functionally or taxonomically characterized. It is for this reason that efforts remain to assemble metatranscriptomic data (together with metagenomic data from the same, or similar sample, if available).

While metatranscriptomic data deposited into public repositories enable future big data analytics and global meta-analyses for discovery of important genes, pathways, and organisms, a prerequisite is the concomitant availability of sample and experimental metadata that help define the context of these complex datasets. While over time, a larger fraction of available metatranscriptomes has been deposited with some metadata (**Figure 1**), to realize the full potential of metatranscriptomic meta-analyses, or for any form of metatranscriptome reanalysis, the deposition of adequate sample metadata should become an important focus of future efforts, together with standardization of vocabulary for metadata descriptors. Several grass-roots efforts among the larger scientific community such as Minimum Information about any Sequence or MIxS (Yilmaz et al., 2011) will be needed if we hope to set a series of standards for inclusion of sufficiently detailed metadata when depositing metatranscriptomic (or any omics) datasets that would allow such all-inclusive analyses.

Because of the broad dynamic range of both microbiome membership relative abundance and of gene expression within any given organism, metatranscriptomics requires a very large number of data points (i.e. reads). Therefore, high throughput short read technologies dominate this area, however the rise of long read technologies holds great promise when throughput (per dollar) improves. Longer reads will be able to help with all aspects of analysis (assembly, taxonomy determination, functional analysis), and will additionally provide better resolution of transcript isoforms, polycistronic operons, and different genes with high similarity.

While today's studies are primarily performed with a single short read technology (i.e. Illumina), there exist a large number of analytical tools to aid in all aspects of data analysis. In this review, we highlight some of the major methods of analyzing metatranscriptomics data, some of the specific bioinformatics tools used to accomplish these analyses, and some more complex metatranscriptomic workflows that combine a number of these tools to address several biological questions with minimal input or effort from the users. Each of the workflows uses either a read-based or an assembly-based approach towards taxonomic and/or functional analysis of organisms and genes expressed within a community, and their relative abundances. Some of the workflows can even proceed all the way to performing differential gene expression analysis among various input samples. While the workflows share a number of similarities, the tools used differ, and it is not clear which workflow, or bioinformatics tool, may be best under any given scenario. Thus, one additional area that beckons for more research is the benchmarking of the performance and accuracy of bioinformatics tools and pipelines with metatranscriptomic data. The complexity of real microbiomes and our incomplete knowledge of the organisms (or genome sequences) present within them have been great challenges in trying to perform such benchmarking experiments. While we have yet to create tools that are truly able to mimic real sequencing datasets, methods that generate simulated sequencing data from known genomes may be used to create a range of simulated metatranscriptome datasets that can in turn be used to test the behavior of bioinformatics tools and parameter settings. Past efforts have focused on *ad hoc* metrics to evaluate performance using real samples and sequencing data. To make matters more complex, further advancements in sequencing technologies will continue to push the development of new tools and workflows. An accepted framework for benchmarking new tools would help the field progress, and possibly coalesce towards accurate and appropriate workflows. Despite some of the issues with metatranscriptomics as a method, the continued development of new tools and algorithms for analyzing metatranscriptomic data coupled with our increasing understanding of the challenges presented by such datasets, it is clear that the next generation of metatranscriptomics tools hold great promise in facilitating our understanding of the biologically active fraction of microbiomes, and the relevant pathways involved.

AUTHOR CONTRIBUTIONS

PC and MS wrote the manuscript with inputs from CL. All authors read and approved the manuscript.

FUNDING

This work was supported by the US Defense Threat Reduction Agency (grants CB10152, R-00480-16-0, and CB10623 to PC) and by the U.S Department of Energy, Office of Science, Biological and Environmental Research Division, under award numbers LANLF59T and LANLF59C to PC." (KP1601010 and 4000150817 877 to PC).

ACKNOWLEDGMENTS

We would like to thank Yan Xu for providing metatranscriptomic data from SRA database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00904/full#supplementary-material

REFERENCES

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Bao, G., Wang, M., Doak, T. G., and Ye, Y. (2015). Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. *Front. Microbiol.* 6, 896. doi: 10.3389/fmicb.2015.00896

Bashiardes, S., Zilberman-Schapira, G., and Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* 10, 19–25. doi: 10.4137/BBI.S34610

Berg, C., Dupont, C. L., Asplund-Samuelsson, J., Celepli, N. A., Eiler, A., Allen, A. E., et al. (2018). Dissection of microbial community functions during a cyanobacterial bloom in the baltic sea *via* metatranscriptomics. *Front. Mar. Sci.* 5, 55. doi: 10.3389/fmars.2018.00055

Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X., et al. (2015). Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* 13, 390–401. doi: 10.1016/j.csbj.2015.06.001

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

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

Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A. D. (2019). rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience* 8. doi: 10.1093/gigascience/gi2100

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

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Crump, B. C., Wojahn, J. M., Tomas, F., and Mueller, R. S. (2018). Metatranscriptomics and amplicon sequencing reveal mutualisms in seagrass microbiomes. *Front. Microbiol.* 9, 388. doi: 10.3389/fmicb.2018.00388

Damon, C., Lehembre, F., Oger-Desfeux, C., Luis, P., Ranger, J., Fraissinet-Tachet, L., et al. (2012). Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PLoS One* 7, e28967. doi: 10.1371/journal.pone.0028967

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

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S .A.* 111, E2329–E2338. doi: 10.1073/pnas.1319284111

Freitas, T.A.K., Li, P.-E., Scholz, M. B., and Chain, P. S. G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* 43, e69. doi: 10.1093/nar/gkv180

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., et al. (2008). Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3805–3810. doi: 10.1073/pnas.0708897105

Geisen, S., Tveit, A. T., Clark, I. M., Richter, A., Svenning, M. M., Bonkowski, M., et al. (2015). Metatranscriptomic census of active protists in soils. *ISME J.* 9, 2178–2190. doi: 10.1038/ismej.2015.30

Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., et al. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* 3, e3042. doi: 10.1371/journal.pone.0003042

Gonzalez, E., Pitre, F. E., Pagé, A. P., Marleau, J., Guidi Nissim, W., St-Arnaud, M., et al. (2018). Trees, fungi and bacteria: tripartite metatranscriptomics of a root microbiome responding to soil contamination. *Microbiome* 6, 53. doi: 10.1186/s40168-018-0432-5

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

Hatton, J. J., Stevenson, T. J., Buck, C. L., and Duddleston, K. N. (2017). Diet affects arctic ground squirrel gut microbial metatranscriptome independent of community structure. *Environ. Microbiol.* 19, 1518–1535. doi: 10.1111/1462-2920.13712

Hayden, H. L., Savin, K. W., Wadeson, J., Gupta, V. V. S. R., and Mele, P. M. (2018). Comparative metatranscriptomics of wheat rhizosphere microbiomes in disease suppressive and non-suppressive soils for *Rhizoctonia solani* AG8. *Front. Microbiol.* 9, 859. doi: 10.3389/fmicb.2018.00859

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

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

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

Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016a). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26, 1721–1729. doi: 10.1101/gr.210641.116

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., and Zhan, X. (2016b). FMAP: Functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinf.* 17, 420. doi: 10.1186/s12859-016-1278-0

Klingenberg, H., and Meinicke, P. (2017). How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ* 5, e3859. doi: 10.7717/peerj.3859

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Leung, H. C. M., Yiu, S.-M., Parkinson, J., and Chin, F. Y. L. (2013). IDBA-MT: *de novo* assembler for metatranscriptomic data generated from next-generation sequencing technology. *J. Comput. Biol.* 20, 540–550. doi: 10.1089/cmb.2013.0042

Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2014). IDBA-MTP: a hybrid metatranscriptomic assembler based on protein information. *Res. Comput. Mol. Biol.* 22(5). doi: 10.1007/978-3-319-05269-4_12

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, F., Hitch, T. C. A., Chen, Y., Creevey, C. J., and Guan, L. L. (2019). Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. *Microbiome* 7, 6. doi: 10.1186/s40168-019-0618-5

Li, P.-E., Lo, C.-C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* 45, 67–80. doi: 10.1093/nar/gkw1027

Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., Mcmullen, R., et al. (2016). BinPacker: packing-based *de novo* transcriptome assembly from RNA-seq data. *PLoS Comput. Biol.* 12, e1004772. doi: 10.1371/journal.pcbi.1004772

Lo, C.-C., and Chain, P. S. G. (2014). Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinf.* 15, 366. doi: 10.1186/s12859-014-0366-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinf.* 10, 161. doi: 10.1186/1471-2105-10-161

Mann, E., Wetzels, S. U., Wagner, M., Zebeli, Q., and Schmitz-Esser, S. (2018). Metatranscriptome Sequencing Reveals Insights into the Gene Expression and Functional Potential of Rumen Wall Bacteria. *Front. Microbiol.* 9, 43. doi: 10.3389/fmicb.2018.00043

Marcelino, V. R., Wille, M., Hurt, A. C., Gonzalez-Acuna, D., Klaassen, M., Schlub, T. E., et al. (2019). Meta-transcriptomics reveals a diverse antibiotic resistance gene pool in avian microbiomes. *BMC Biol.* 17, 31. doi: 10.1186/s12915-019-0649-1

Martinez, X., Pozuelo, M., Pascal, V., Campos, D., Gut, I., Gut, M., et al. (2016). MetaTrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.* 6, 26447. doi: 10.1038/srep26447

Meinicke, P. (2015). UProC: tools for ultra-fast protein domain classification. *Bioinformatics* 31, 1382–1388. doi: 10.1093/bioinformatics/btu843

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7, 11257. doi: 10.1038/ncomms11257

Moniruzzaman, M., Wurch, L. L., Alexander, H., Dyhrman, S. T., Gobler, C. J., and Wilhelm, S. W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* 8, 16054. doi: 10.1038/ncomms16054

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

Neves, A. L. A., Li, F., Ghoshal, B., Mcallister, T., and Guan, L. L. (2017). Enhancing the resolution of rumen microbial classification from metatranscriptomic data using Kraken and Mothur. *Front. Microbiol.* 8, 2445. doi: 10.3389/fmicb.2017.02445

Ni, Y., Li, J., and Panagiotou, G. (2016). COMAN: a web server for comprehensive metatranscriptomics analysis. *BMC Genomics* 17, 622. doi: 10.1186/s12864-016-2964-z

Nowicki, E. M., Shroff, R., Singleton, J. A., Renaud, D. E., Wallace, D., Drury, J., et al. (2018). Microbiota and metatranscriptome changes accompanying the onset of gingivitis. *MBio* 9, 1–17. doi: 10.1128/mBio.00575-18

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1101/gr.213959.116

O'leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., Mcveigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Parro, V., Moreno-Paz, M., and González-Toril, E. (2007). Analysis of environmental transcriptomes by DNA microarrays. *Environ. Microbiol.* 9, 453–464. doi: 10.1111/j.1462-2920.2006.01162.x

Peng, Y., Leung, H. C. M., Yiu, S.-M., Lv, M.-J., Zhu, X.-G., and Chin, F. Y. L. (2013). IDBA-tran: a more robust *de novo* de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* 29, i326–i334. doi: 10.1093/bioinformatics/btt219

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Pérez-Losada, M., Castro-Nallar, E., Bendall, M. L., Freishtat, R. J., and Crandall, K. A. (2015). Dual transcriptomic profiling of host and microbiota during health and disease in pediatric asthma. *PLoS One* 10, e0131819. doi: 10.1371/journal.pone.0131819

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. (2018). Long reads: their purpose and place. *Hum. Mol. Genet.* 27, R234–R241. doi: 10.1093/hmg/ddy177

Poretsky, R. S., Bano, N., Buchan, A., Lecleir, G., Kleikemper, J., Pickering, M., et al. (2005). Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* 71, 4121–4126. doi: 10.1128/AEM.71.7.4121-4126.2005

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

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

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). *De novo* assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi: 10.1038/nmeth.1517

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

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., et al. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541–2547. doi: 10.1128/AEM.66.6.2541-2547.2000

Rotmistrovsky, K., and Agarwala, R. (2011). BMTagger: best match tagger for removing human reads from metagenomics datasets.

Saminathan, T., Garcia, M., Ghimire, B., Lopez, C., Bodunrin, A., Nimmakayala, P., et al. (2018). Metagenomic and metatranscriptomic analyses of diverse watermelon cultivars reveal the role of fruit associated microbiome in carbohydrate metabolism and ripening of mature fruits. *Front. Plant Sci.* 9, 4. doi: 10.3389/fpls.2018.00004

Schirmer, M., Franzosa, E. A., Lloyd-Price, J., Mciver, L. J., Schwager, R., Poon, T. W., et al. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat. Microbiol.* 3, 337–346. doi: 10.1038/s41564-017-0089-z

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086–1092. doi: 10.1093/bioinformatics/bts094

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

Sequeira, J. C., Rocha, M., Madalena Alves, M., and Salvador, A. F. (2019). "MOSCA: an automated pipeline for integrated metagenomics and metatranscriptomics data analysis," in *Practical Applications of Computational Biology and Bioinformatics, 12th International Conference* (Springer International Publishing). doi: 10.1007/978-3-319-98702-6_22

Shi, Y., Tyson, G. W., and Delong, E. F. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 459, 266–269. doi: 10.1038/nature08055

Sollinger, A., Tveit, A. T., Poulsen, M., Noel, S. J., Bengtsson, M., Bernhardt, J., et al. (2018). Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. *mSystems* 3, 1–19. doi: 10.1128/mSystems.00038-18

Tamames, J., and Puente-Sánchez, F. (2018). SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* 9, 3349. doi: 10.3389/fmicb.2018.03349

Tarazona, S., Furio-Tari, P., Turra, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* 43, e140. doi: 10.1093/nar/gkv711

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

Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., et al. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME J.* 7, 2248–2258. doi: 10.1038/ismej.2013.119

Ugarte, A., Vicedomini, R., Bernardes, J., and Carbone, A. (2018). A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome* 6, 149. doi: 10.1186/s40168-018-0532-2

Uniprot, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., and Lemay, D. G. (2018). SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinf.* 19, 175. doi: 10.1186/s12859-018-2189-z

White, R. A., 3rd, Bottos, E. M., Roy Chowdhury, T., Zucker, J. D., Brislawn, C. J., Nicora, C. D., et al. (2016). Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. *mSystems* 1, 1–15. doi: 10.1128/mSystems.00045-16

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., et al. (2016). The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* 44, D590–D594. doi: 10.1093/nar/gkv1322

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

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30, 1660–1666. doi: 10.1093/bioinformatics/btu077

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011). iPath2.0: interactive pathway explorer. *Nucleic Acids Res.* 39, W412–W415. doi: 10.1093/nar/gkr313

Ye, Y., and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* 5, e1000465. doi: 10.1371/journal.pcbi.1000465

Ye, Y., and Tang, H. (2016). Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics* 32, 1001–1008. doi: 10.1093/bioinformatics/btv510

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415. doi: 10.1038/nbt.1823

Zhong, C., Edlund, A., Yang, Y., Mclean, J. S., and Yooseph, S. (2016). Metagenome and metatranscriptome analyses using protein family profiles. *PLoS Comput. Biol.* 12, e1004991. doi: 10.1371/journal.pcbi.1004991

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

frontiers in Genetics

METHODS
published: 15 October 2019
doi: 10.3389/fgene.2019.00999

An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs

Xiaoli Dong and Marc Strous*

Department of Geoscience, University of Calgary, Calgary, AB, Canada

Here, we describe MetaErg, a standalone and fully automated metagenome and metaproteome annotation pipeline. Annotation of metagenomes is challenging. First, metagenomes contain sequence data of many organisms from all domains of life. Second, many of these are from understudied lineages, encoding genes with low similarity to experimentally validated reference genes. Third, assembly and binning are not perfect, sometimes resulting in artifactual hybrid contigs or genomes. To address these challenges, MetaErg provides graphical summaries of annotation outcomes, both for the complete metagenome and for individual metagenome-assembled genomes (MAGs). It performs a comprehensive annotation of each gene, including taxonomic classification, enabling functional inferences despite low similarity to reference genes, as well as detection of potential assembly or binning artifacts. When provided with metaproteome information, it visualizes gene and pathway activity using sequencing coverage and proteomic spectral counts, respectively. For visualization, MetaErg provides an HTML interface, bringing all annotation results together, and producing sortable and searchable tables, collapsible trees, and other graphic representations enabling intuitive navigation of complex data. MetaErg, implemented in Perl, HTML, and JavaScript, is a fully open source application, distributed under Academic Free License at https://github.com/xiaoli-dong/metaerg. MetaErg is also available as a docker image at https://hub.docker.com/r/xiaolidong/docker-metaerg.

OPEN ACCESS

Edited by:
Bas E. Dutilh,
Utrecht University,
Netherlands

Reviewed by:
Cuncong Zhong,
University of Kansas,
United States
João Marcelo Pereira Alves,
University of São Paulo,
Brazil
Julien Tremblay,
National Research Council Canada
(NRC-CNRC), Canada

****Correspondence:***
Xiaoli Dong
xdong@ucalgary.ca

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

Received: *08 April 2019*
Accepted: *20 September 2019*
Published: *15 October 2019*

Citation:
Dong X and Strous M (2019) An
Integrated Pipeline for
Annotation and Visualization of
Metagenomic Contigs.
Front. Genet. 10:999.
doi: 10.3389/fgene.2019.00999

Keywords: metagenomics, metaproteomics, bioinformatics, gene prediction, functional annotation, taxonomic classification, pathway prediction, visualization

INTRODUCTION

Genome annotation is, literally, the annotation of features on assembled DNA molecules. Such features are, in the first place, genes, including those encoding proteins ["open reading frames" (ORFs)] and those encoding ribosomal or transfer RNA molecules. Annotation consists of the identification of such features and providing each feature with a meaningful list of hints about its possible biological function. Annotation is usually the final step of the automated computational processing of genomic or metagenomic data and the beginning of biology.

Depending on their background and research question, biologists will have different annotation needs. For example, when the research targets a single microbe, detailed gene-by-gene annotation of its genome would be desired. On the other hand, when the research targets a complete ecosystem, a high level summary of the functional potential of the associated metagenome might be the aim. These examples also display a different starting point for annotation. In the first case, it may consist of a single, near-perfect whole genome sequence. In the second case, it may consist of many MAGs of varying quality, unbinned metagenomic contigs, or even billions of unassembled reads.

What sets annotation apart from other computational steps in processing metagenomic data is that no benchmarks for annotation tools exist. That means that ranking these tools and objectively declaring a winner is not straightforward. The choice of the best annotation pipeline will depend on the data, the research question, the computational resources available, and the background of the researcher who needs to make sense of the annotation software's hints and the way they are presented.

In practice, options for genome annotation come in two flavors: online platforms and standalone pipelines. Examples of online platforms are IMG (Chen et al., 2017), MG-RAST (Keegan et al., 2016), MicroScope (Vallenet et al., 2017), Mgnify (Mitchell et al., 2018), and Edge (Li et al., 2017). When opting for a platform, you avoid the need for local computational infrastructure or tedious installation and updating of tools and databases, while benefiting from online collaboration abilities. The platform may provide accession numbers for sharing data after publication, as these platforms may also be data repositories.

However, a platform might not offer the type of annotation needed for a specific research question or might be slower in the uptake of the latest selection of tools and databases. If such factors are important, opting for a standalone pipeline might be the way to go. Scientists who are fluent in scripting languages, such as Python or Perl, might even create their own pipeline from scratch. Examples of available standalone pipelines for annotation of assembled contigs, scaffolds, or whole genome sequences are Prokka (Seemann, 2014), DFAST-core (Tanizawa et al., 2018), and PGAP (Tatusova et al., 2016). Prokka is a very fast genome annotation pipeline. Its core concept is that some databases or tools provide better or more information than others. Once a gene is annotated with a positive "hit" to a good database, there is no need to perform additional searches. DFAST adds to this approach by using a faster similarity search tool (ghostx). It infers orthology assignments based on reciprocal-best-blast-hits between the query genome and a larger set of reference genomes, potentially including user-added custom reference genomes. It is thus especially useful to transfer annotations from a well-annotated reference genome. PGAP is used by the NCBI to annotate submitted whole genome sequences. It combines sophisticated gene prediction algorithms with gene assignments to its set of prokaryotic protein clusters (Klimke et al., 2009). As an institutional "gold standard" annotation, it emphasizes annotation standards and conventions, quality control, and due diligence during execution.

Here, we present MetaErg, an extendable standalone annotation pipeline developed for metagenome-assembled genomes (MAGs). Genome-centric metagenome data provides three major challenges. The first is that assembly quality can be relatively poor, and some contamination of MAGs with "foreign" genes can be expected. This challenge is addressed by performing fast similarity searches against a much larger database than would be needed to simply infer functions, to classify each gene taxonomically. This enables detection of potentially artefactual, hybrid bins or contigs. The second is that the user will likely need to make sense of many annotated genomes simultaneously. This challenge is addressed by visualizing and summarizing data, to enable quick inferences about encoded biological functions and pathways. The third is that, for many environmental microorganisms, meaningful/close reference genomes are not yet available. This challenge is addressed by always providing comprehensive information about each gene, derived from different tools and databases, to assign functions as well as practically possible for genes with low similarity to reference genes.

MATERIALS AND METHODS

Program Implementation Overview

MetaErg is an integrated and fully automated pipeline for annotating metagenome-assembled contigs. It integrates a number of open-source tools and its modular design allows for flexible workflows, addition of new functions, and easy refactoring. MetaErg's implementation consists of five main modules (**Figure 1**), including a command-line interface, an input data preprocessing module for filtering and formatting input DNA sequences, a structural annotation module for predicting biological features and elements, a function annotation module for inferring gene functions and classifying rRNA genes and ORFs to taxonomic lineages, and a presentation module for presenting annotation results in various summary reports and for visualization using HTML and JavaScript.

Command Line Interface

MetaErg is a command line program, designed to run on a Linux server or cluster. It accepts a preassembled FASTA format DNA sequence file as the minimum required input. The default values for the optional parameters in the pipeline are optimized for metagenome analysis. Through a command-line interface, experienced users can interact with the program to customize the gene prediction and database searching parameters, enable or disable certain tools and functions, setup data filtering thresholds, and specify an output directory.

Sequence Data Preprocessing

Every input DNA sequence is inspected, validated, and reformatted before annotation. The sequence identifiers in the input file must be unique; otherwise, the input file will be rejected, and the annotation process will be terminated. Any ambiguous nucleotides in the input sequence file are replaced by N. Gaps (-) and pads (*) are removed. Sequences shorter than a user defined minimum length are removed.

Structural Annotation

MetaErg begins biological feature and element prediction by identifying CRISPR elements and noncoding RNA genes (tRNA, rRNA genes). Next, to avoid identification of artefactual protein

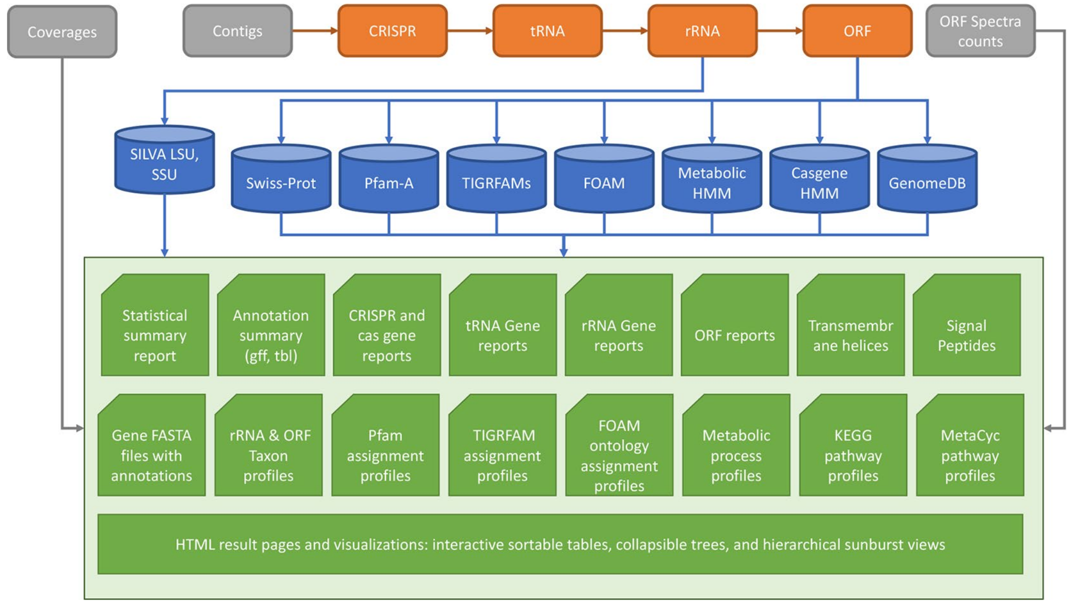

Coverages
Contigs
CRISPR
tRNA
rRNA
ORF
ORF Spectra counts
SILVA LSU, SSU
Swiss-Prot
Pfam-A
TIGRFAMs
FOAM
Metabolic HMM
Casgene HMM
GenomeDB
Statistical summary report
Annotation summary (gff, tbl)
CRISPR and cas gene reports
tRNA Gene reports
rRNA Gene reports
ORF reports
Transmembrane helices
Signal Peptides
Gene FASTA files with annotations
rRNA & ORF Taxon profiles
Pfam assignment profiles
TIGRFAM assignment profiles
FOAM ontology assignment profiles
Metabolic process profiles
KEGG pathway profiles
MetaCyc pathway profiles
HTML result pages and visualizations: interactive sortable tables, collapsible trees, and hierarchical sunburst views

FIGURE 1 | MetaErg annotation workflow. The input file to MetaErg is a FASTA file that contains the assembled contigs.

coding genes overlapping with detected noncoding features, MetaErg masks these features by replacing them with Ns. Next, protein encoding genes are predicted. **Figure 1** shows the MetaErg workflow.

The identification of CRISPR elements is achieved using MinCED (Skennerton, 2016) with default parameters. tRNA genes are predicted with the ARAGORN program (Laslett and Canback, 2004).

Ribosomal RNA genes (5S, 5.8S, 16S, 18S, 23S, 28S) are identified and classified using rRNAFinder, an in-house tool package, which is included in the MetaErg release. rRNAFinder uses nhmmer (Wheeler and Eddy, 2013) to query locally built rRNA HMM profiles against the input contig sequences for detecting rRNA genes on the contigs. To build the rRNA HMM profiles, the "rfam.seed.gz" file was downloaded from the Rfam database (Kalvari et al., 2018). The FASTA-formatted rRNA gene alignments were extracted and written to separate files for each of the three domains of life (*Bacteria, Archaea, Eukaryota*), respectively. The alignment files for each domain were then used by the hmmbuild program in HMMER (Eddy, 2011) to create an rRNA gene HMM profile for the domain. Because a metagenome may contain rRNA sequences from all domains of life, in "metagenome" mode, rRNAFinder uses HMM models from all three domains of life. When multiple models yield hits to the same region, rRNAFinder outputs only the result of the model with the lowest *E*-value. When the *E*-value is the same for multiple hits, all best scoring predictions are kept. rRNAFinder uses blastn (Altschul et al., 1990) for classification of detected rRNA genes using the full-length SILVA SSU and LSU database (Quast et al., 2012). The standalone rRNAFinder tool is also freely available at https://github.com/xiaoli-dong/rRNAFinder.

Protein coding genes (ORFs) are predicted using Prodigal (Hyatt et al., 2010). ORFs shorter than 180 nucleotides are excluded from further analysis by default.

Functional and Taxonomic Annotation

Metagenome functional annotation is very similar to genomic annotation and relies on comparisons of predicted genes with existing, previously annotated sequences. The goal is to propagate accurate annotations to correctly identified orthologs (Kunin et al., 2008).

Firstly, predicted ORFs are run through motif prediction tools. SignalP 5.0 (Armenteros et al., 2019) is run on all ORFs to predict the presence and absence of signal peptides and the location of their cleavage sites within an ORF. TMHMM (Krogh et al., 2001) is run on all ORFs to detect the transmembrane helices.

MetaErg uses profile HMMs and blast-based searches to detect similarity. All ORFs are searched against different databases. All search results are combined to associate query genes with functional categories, protein domains, KEGG Orthology (KO) terms, Gene Ontology (GO) terms, Enzyme Commission (EC) numbers, and metabolic potentials and traits. In brief, ORFs are searched with HMMs from Pfam-A (Finn et al., 2014), TIGRFAM (Haft et al., 2013), FOAM (Prestat et al., 2014), Metabolic-hmm (Anantharaman et al., 2016), and casgenes.hmm (Burstein et al., 2016) using the hmmsearch tool. ORFs are also searched against the SwissProt (BBairoch and Apweiler, 2000) database using DIAMOND (Buchfink et al., 2014). ORFs without any search outcomes are annotated as "hypothetical protein".

MinPath (Minimal set of Pathways) was used to reconstruct metabolic pathways. MinPath minimizes parsimony and yields a conservative estimate of the biological pathways present in a query dataset (Ye and Doak, 2009). MetaErg uses MinPath to predict KEGG (Kanehisa and Goto, 2000) and MetaCyc (Karp et al., 2002) pathways. For predicting the minimal set of KEGG pathways that still explains the presence of the detected functional genes, an ORF-identifier-to-KO-number-mapping-file is provided as the input to MinPath. For inferring the list of MetaCyc pathways, an ORF-identifier-to-EC-number-mapping-file is provided as the input to MinPath. The mapping files are derived from the blast searches of the ORFs against the Swiss-Prot databases, as well as HMM searches against the FOAM and the TIGRFAMs database.

MetaErg classifies all ORFs based on best DIAMOND hits against a custom database, GenomeDB. To build GenomeDB, the Genome Taxonomy Database (GTDB) gtdbtk_r89_data.tar.gz (Parks et al., 2018) was downloaded from https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/. Each genome included in GTDB was checked for presence in the NCBI RefSeq database. If present, the FASTA-formatted protein files were downloaded. Otherwise, the ORFs for the genome were predicted using Prodigal. The downloaded and locally predicted ORFs inherited their taxonomy from GTDB. To the GTDB data, only associated with *Bacteria* and *Archaea*, we added *Eukaryota* and viral data, by downloading the available NCBI RefSeq protein sequences of unicellular protozoa, fungi, plants (excluding *Embryophyta*), and viruses. The taxonomy of those proteins in GenomeDB was inherited from the NCBI records. For that, we inspected the assembly_summary.txt file, present in each NCBI RefSeq subdirectory (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/), which associates each assembled genome with a "ftp_path" and a "species_taxid". We retrieved the protein sequences of each available Eukaryote or viral genome by following "ftp_path". The taxonomy of the protein sequences was obtained *via* "species_taxid". This process was automated in a Perl script, enabling periodical updating of the database.

With a user-supplied coverage file generated by mapping reads from each sample to the assembled contig sequences, MetaErg quantifies the relative abundance of organisms, functions, metabolic processes, and pathways in each sample by tracking the number of reads that map to each gene family or orthologous group. The coverage file, generated using "jgi_summarize_bam_contig_depths" from MetaBat (Kang et al., 2015), is a tab delimited text file and the example coverage file is available at https://github.com/xiaoli-dong/metaerg/blob/master/example/demo.depth.txt. With a user-supplied metaproteomics spectral count file, MetaErg quantifies the abundance (in the proteome) of each taxon, function, metabolic process, and pathway based on expressed genes included in the spectral counts file for each sample. The spectral count file is a tab-delimited text file. The first column of the file is the gene id and all the columns after are the normalized protein expression level. The example metaproteomics spectral count file is available at https://github.com/xiaoli-dong/metaerg/blob/master/example/demo.plevel.txt.

Output and Visualization

MetaErg reports annotations at the individual gene, single genome, and community level. For each gene, it reports the taxonomic classification and functional annotations, GO terms, EC numbers, KO terms, and its association with a metabolic pathway. At the community or genome level, MetaErg presents the taxonomic composition, protein function profiles, metabolic process profiles, and metabolic pathway profiles. A MetaErg output demo page is available at https://xiaoli-dong.github.io/metaerg/

To facilitate the exploration of complex metagenome annotation results and make sense of the data, MetaErg's annotation reports are presented in various formats. The HTML result page (**Figure 2**) visually brings together text summaries, output data files, and accompanying visualizations. The interactive sortable and searchable gene, function, and profile tables, collapsible trees, sunburst hierarchical views of taxonomy and functional ontology, and other graphical representations, enable the effective interactive exploration, analysis, filtering, and intuitive navigation of complex metagenomic data (**Figure 3**).

The intermediate results, including those from feature predictions and similarity searches, are stored as files, which could be used to dig deeper into the data and validate the results later on. With the intermediate files in place, MetaErg will skip the steps used to generate them when the program is restarted with the same input parameters. This can greatly reduce the computational time when redoing the analysis.

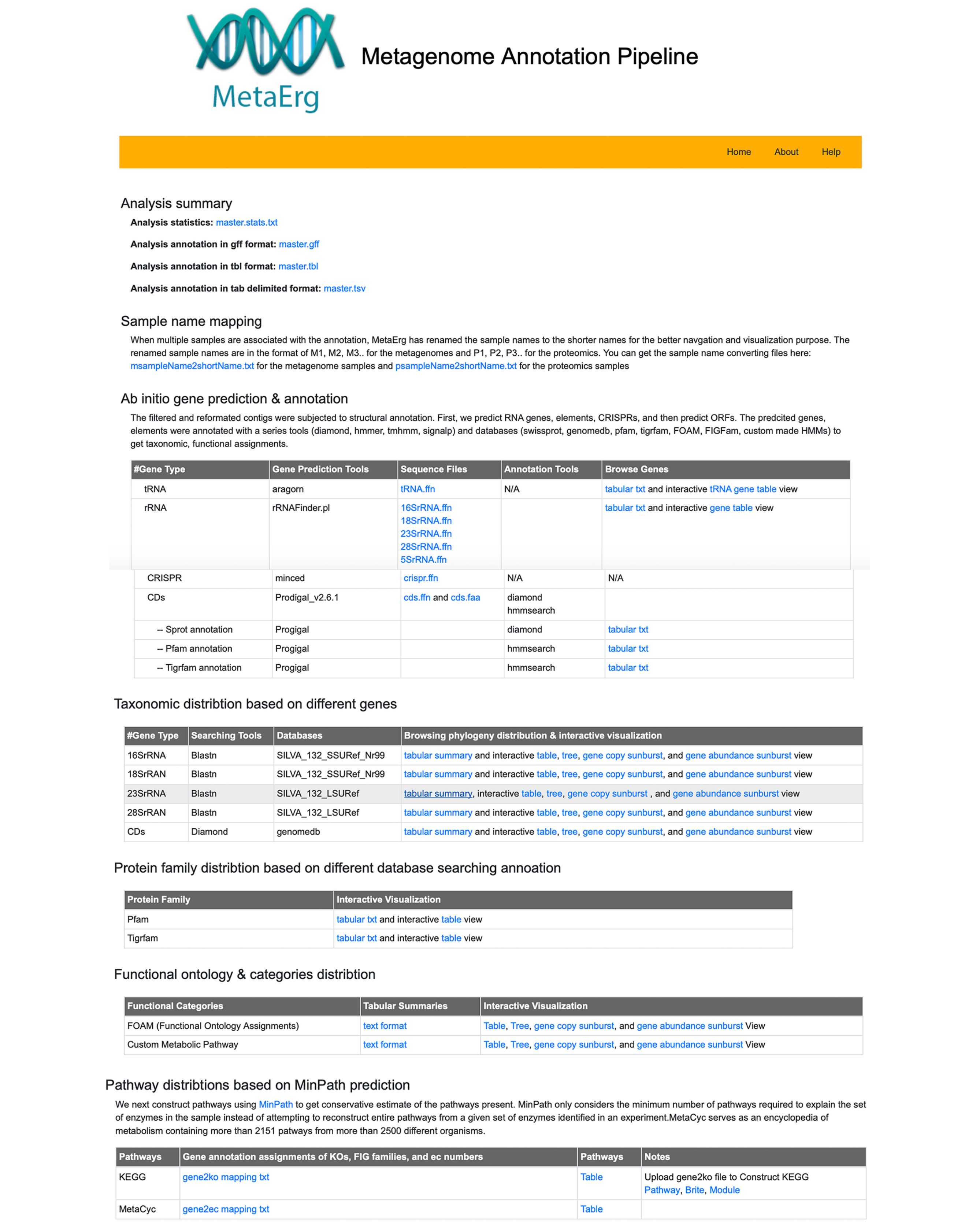
Generation of the Test Dataset

The paired-end Illunima raw reads of three biological replicates of a mock community sample (Kleiner et al., 2017, NCBI SRA accession numbers ERR1877474, ERR1877474, and ERR1877476) were filtered using BBDuk from the BBTools suite (Bushnell, 2014). Briefly, each read was screened by reference and by kmer for Illumina adapters (options: tbo tpe k = 23 mink = 11 hdist = 1 ktrim = r) and for Phix (options: k = 31 hdist = 1) and quality trimmed and filtered (options: qtrim = rl trimq = 15 minlength = 30 entropy = 0.5). After cleaning, the remaining reads were merged using BBMerge with default settings. The resulting merged single-end reads and unmerged paired-end reads from three samples were co-assembled together using metaSpades (Nurk et al., 2017) with default settings. After assembling, contigs shorter than 500 bp were excluded from further analysis.

Mapping of the quality-controlled reads from all three libraries back to the assembled contigs was preformed using BBMap with default settings. The depth coverage file "depth.txt" was generated using "jgi_summarize_bam_contig_depths" from MetaBat.

RESULTS

To test MetaErg and determine the computational footprint, a MetaErg job was submitted to a Linux cluster node (56 threads, 256 GB RAM) with the assembled contigs from a mock community as the input. The mock community consisted of 25 species of *Bacteria*, 1 *Archaeon*, 1 *Eukaryote*, and 5 *phages* (Kleiner et al., 2017). Assembly with MetaSpades resulted in 4,576 contigs (N50 126,358 base pairs, 85,113,339 base pairs



Metagenome Annotation Pipeline

Analysis summary

Analysis statistics: master.stats.txt

Analysis annotation in gff format: master.gff

Analysis annotation in tbl format: master.tbl

Analysis annotation in tab delimited format: master.tsv

Sample name mapping

When multiple samples are associated with the annotation, MetaErg has renamed the sample names to the shorter names for the better navgation and visualization purpose. The renamed sample names are in the format of M1, M2, M3.. for the metagenomes and P1, P2, P3.. for the proteomics. You can get the sample name converting files here: msampleName2shortName.txt for the metagenome samples and psampleName2shortName.txt for the proteomics samples

Ab initio gene prediction & annotation

The filtered and reformated contigs were subjected to structural annotation. First, we predict RNA genes, elements, CRISPRs, and then predict ORFs. The predcited genes, elements were annotated with a series tools (diamond, hmmer, tmhmm, signalp) and databases (swissprot, genomedb, pfam, tigrfam, FOAM, FIGFam, custom made HMMs) to get taxonomic, functional assignments.

#Gene Type	Gene Prediction Tools	Sequence Files	Annotation Tools	Browse Genes
tRNA	aragorn	tRNA.ffn	N/A	tabular txt and interactive tRNA gene table view
rRNA	rRNAFinder.pl	16SrRNA.ffn 18SrRNA.ffn 23SrRNA.ffn 28SrRNA.ffn 5SrRNA.ffn		tabular txt and interactive gene table view
CRISPR	minced	crispr.ffn	N/A	N/A
CDs	Prodigal_v2.6.1	cds.ffn and cds.faa	diamond hmmsearch	
-- Sprot annotation	Progigal		diamond	tabular txt
-- Pfam annotation	Progigal		hmmsearch	tabular txt
-- Tigrfam annotation	Progigal		hmmsearch	tabular txt

Taxonomic distribtion based on different genes

#Gene Type	Searching Tools	Databases	Browsing phylogeny distribution & interactive visualization
16SrRNA	Blastn	SILVA_132_SSURef_Nr99	tabular summary and interactive table, tree, gene copy sunburst, and gene abundance sunburst view
18SrRAN	Blastn	SILVA_132_SSURef_Nr99	tabular summary and interactive table, tree, gene copy sunburst, and gene abundance sunburst view
23SrRNA	Blastn	SILVA_132_LSURef	tabular summary, interactive table, tree, gene copy sunburst , and gene abundance sunburst view
28SrRAN	Blastn	SILVA_132_LSURef	tabular summary and interactive table, tree, gene copy sunburst, and gene abundance sunburst view
CDs	Diamond	genomedb	tabular summary and interactive table, tree, gene copy sunburst, and gene abundance sunburst view

Protein family distribtion based on different database searching annoation

Protein Family	Interactive Visualization
Pfam	tabular txt and interactive table view
Tigrfam	tabular txt and interactive table view

Functional ontology & categories distribtion

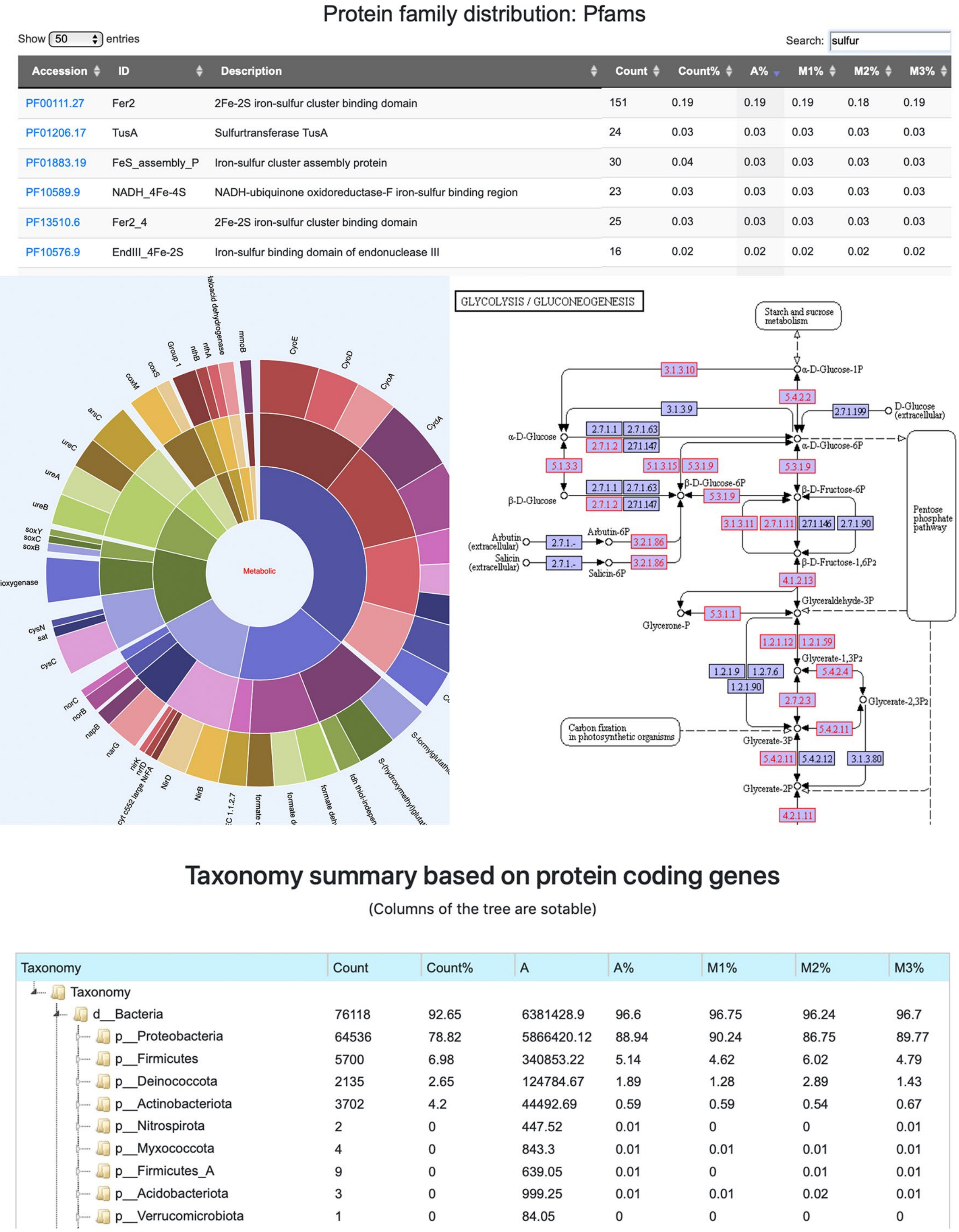
Functional Categories	Tabular Summaries	Interactive Visualization
FOAM (Functional Ontology Assignments)	text format	Table, Tree, gene copy sunburst, and gene abundance sunburst View
Custom Metabolic Pathway	text format	Table, Tree, gene copy sunburst, and gene abundance sunburst View

Pathway distribtions based on MinPath prediction

We next construct pathways using MinPath to get conservative estimate of the pathways present. MinPath only considers the minimum number of pathways required to explain the set of enzymes in the sample instead of attempting to reconstruct entire pathways from a given set of enzymes identified in an experiment.MetaCyc serves as an encyclopedia of metabolism containing more than 2151 patways from more than 2500 different organisms.

Pathways	Gene annotation assignments of KOs, FIG families, and ec numbers	Pathways	Notes
KEGG	gene2ko mapping txt	Table	Upload gene2ko file to Construct KEGG Pathway, Brite, Module
MetaCyc	gene2ec mapping txt	Table	

FIGURE 2 | MetaErg HTML result page visually links extensive analysis text summaries, result data files, and accompanying visualizations together.



Protein family distribution: Pfams

Show 50 entries | Search: sulfur

Accession	ID	Description	Count	Count%	A%	M1%	M2%	M3%
PF00111.27	Fer2	2Fe-2S iron-sulfur cluster binding domain	151	0.19	0.19	0.19	0.18	0.19
PF01206.17	TusA	Sulfurtransferase TusA	24	0.03	0.03	0.03	0.03	0.03
PF01883.19	FeS_assembly_P	Iron-sulfur cluster assembly protein	30	0.04	0.03	0.03	0.03	0.03
PF10589.9	NADH_4Fe-4S	NADH-ubiquinone oxidoreductase-F iron-sulfur binding region	23	0.03	0.03	0.03	0.03	0.03
PF13510.6	Fer2_4	2Fe-2S iron-sulfur cluster binding domain	25	0.03	0.03	0.03	0.03	0.03
PF10576.9	EndIII_4Fe-2S	Iron-sulfur binding domain of endonuclease III	16	0.02	0.02	0.02	0.02	0.02

Taxonomy summary based on protein coding genes

(Columns of the tree are sotable)

Taxonomy	Count	Count%	A	A%	M1%	M2%	M3%
Taxonomy							
d__Bacteria	76118	92.65	6381428.9	96.6	96.75	96.24	96.7
p__Proteobacteria	64536	78.82	5866420.12	88.94	90.24	86.75	89.77
p__Firmicutes	5700	6.98	340853.22	5.14	4.62	6.02	4.79
p__Deinococcota	2135	2.65	124784.67	1.89	1.28	2.89	1.43
p__Actinobacteriota	3702	4.2	44492.69	0.59	0.59	0.54	0.67
p__Nitrospirota	2	0	447.52	0.01	0	0	0.01
p__Myxococcota	4	0	843.3	0.01	0.01	0.01	0.01
p__Firmicutes_A	9	0	639.05	0.01	0	0.01	0.01
p__Acidobacteriota	3	0	999.25	0.01	0.01	0.02	0.01
p__Verrucomicrobiota	1	0	84.05	0	0	0	0

FIGURE 3 | A screenshot montage of MetaErg output showing an example of the interactive Pfam annotation profile table, a hierarchical metabolic process sunburst view, a taxonomic summary tree view, and a KEGG pathway map. In the KEGG pathway map, the KOs presented in the analyzed dataset were highlighted.

total). The MetaErg job took 2.12 h to complete. The total CPU time needed was 50.5 h. When prediction of signal peptides and transmembrane helixes was included (with options "–sp –tm"), the run time and CPU time increased to 3.7 and 56.2 h, respectively. The average memory usage was 3 GB with peaks up to 9.5 GB. The total disk space used for the analysis including the intermediate files was 6.1 GB and the total disk space used for the final results was 482 MB.

The overall metagenome annotation predicted 20 CRISPR arrays, 878 tRNA genes, 70 rRNA genes (16S, 18S, 23S, 28S, 5S, 5.8S rRNA genes), and 80,407 ORFs. Of these, 48,723, 68,578, 22,001, 25,184, 475, and 437 ORFs were annotated with SwissProt, Pfam, TIGRFAM, FOAM, metabolic hmm, and casgene.hmm databases, respectively. Signal peptides were predicted for 1,480 ORFs and transmembrane helices were predicted for 18,766 ORFs. The relative abundances of taxa, functions, and pathways were nearly identical across all three biological replicates of the mock community.

MetaBat binning of the contigs with default parameters produced 14 useful MAGs (>70% completeness, <5% contamination). MetaBat binned relatively few MAGs for this dataset, because the three available read sets were from replicate samples and were not useful for differential coverage based binning. The annotations for each MAG were extracted directly from the overall annotations using MetaErg's utility scripts. The phylogenetic affiliations of MAGs were estimated according to the taxon assignments of ORFs and rRNA genes in the MAGs and visualized in the interactive HTML trees and sunburst hierarchical views. The HTML visualizations can help users visually validate the binning outcomes and identify chimeric MAGs or contamination with genes from other community members (**Figure 4**). Each gene from each MAG was assigned

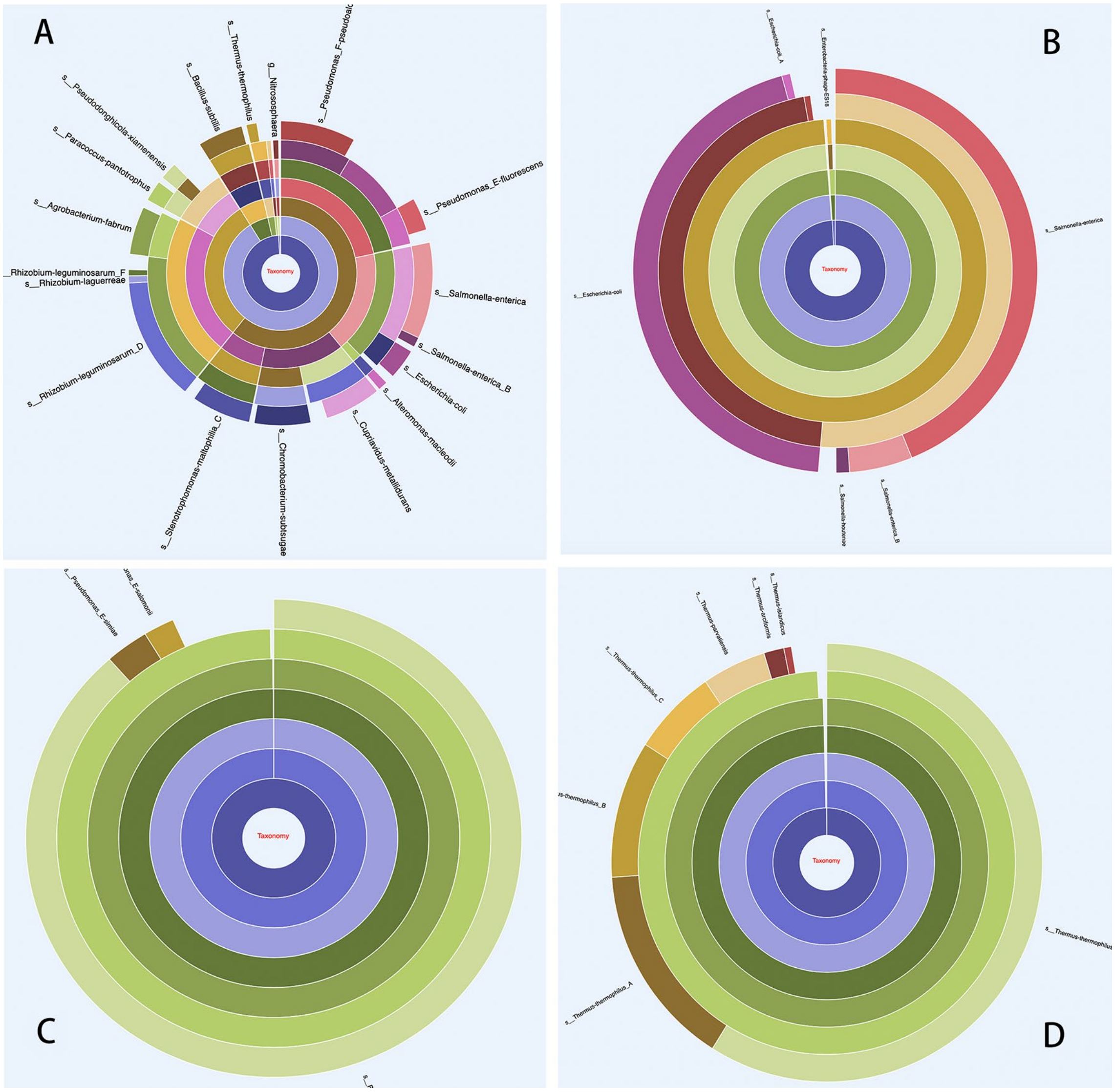


FIGURE 4 | Taxonomy in hierarchical sunburst view. Each taxonomic rank is represented by one ring with the innermost circle representing the kingdom. From the inner to outer rings, the rings represent kingdom, phylum, class, order, family, genus, and species. The segmented areas on the ring are proportional to the relative abundance of the taxon. **(A)** Overall taxonomic distribution profile from all ORFs, which provides insight into the community taxonomic distribution as a whole; **(B)** An example of chimeric MAG, displaying contamination, and this MAG was 99.42% complete and 97.14% contaminated, as assessed by CheckM. The taxon classification profile was based on ORF taxonomic assignment from the MAG; **(C)** and **(D)** Examples of uncontaminated MAGs.

comprehensive information derived from different resources with different tools (**Table 1**).

DISCUSSION

With MetaErg, we provide a standalone and fully automated metagenome and metaproteome annotation pipeline. Compared to other standalone annotation pipelines, such as Prokka (Seemann, 2014) and DFAST-core (Tanizawa et al., 2018), MetaErg requires much more time to run and requires more computational resources. However, these extra resources result in more comprehensive annotation and visualization. Taxonomic classification of each gene, provided by MetaErg, enables detection of potential assembly or binning artifacts, as shown in **Figure 4**. More comprehensive annotation enables better inferences about gene function for genes that are more dissimilar to validated reference genes. High level visualization of pathways, and integration of expression data, enables more effective navigation of the full complexity of a metagenome. Thus, MetaErg provides solutions to challenges specific to metagenomes, which come at a computational cost.

TABLE 1 | An example showing information associated with each protein coding gene after MetaErg analysis.

<table>
<tr><th>TAG</th><th>Value</th></tr>
<tr><td>ID</td><td>mockEvenCell|17112</td></tr>
<tr><td>contigid</td><td>NODE_27_length_371703_cov_24.485093</td></tr>
<tr><td>allec_ids</td><td>7.1.1.-; 1.8.4.8</td></tr>
<tr><td>allko_ids</td><td>K00390; K00338;</td></tr>
<tr><td>allko_ontology</td><td>L1:18_Sulfur compounds metabolism;L2:Sulfur compounds cycle;L3:Sulfate reduction (assimilatory);L4:;</td></tr>
<tr><td>depth</td><td>82.0316;</td></tr>
<tr><td>foam_ecs</td><td>1.8.4.8;</td></tr>
<tr><td>foam_kos</td><td>K00390;</td></tr>
<tr><td>foam_target</td><td>db:FOAM-hmm_rel1a.hmm|HMMsoil748 63 117 evalue:2.5e-13 qcov:30.55 identity:40.00 score:41.9 seqT:47.9 name:KO:K00390_1.8.4.8;</td></tr>
<tr><td>genomedb_oc</td><td>d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Betaproteobacteriales;f__Burkholderiaceae;g__Cupriavidus;</td></tr>
<tr><td>genomedb_target</td><td>db:genomedb|GCA_900185755.1|FYAX01000037.1_317 1 163 evalue:1.4e-89 qcov:100.00 identity:100.00;</td></tr>
<tr><td>pfam_desc</td><td>4Fe-4S binding domain;</td></tr>
<tr><td>pfam_id</td><td>Fer4;</td></tr>
<tr><td>pfam_target</td><td>db:Pfam-A.hmm|PF00037.27 61 80 evalue:2e-07 qcov:12.22 identity:55.00 score:24.1 seqT:53.6 name:Fer4; db:Pfam-A.hmm|PF00037.27 97 118 evalue:5.5e-11 qcov:13.44 identity:63.64 score:35.4 seqT:53.6 name:Fer4;</td></tr>
<tr><td>sport_desc</td><td>NADH-quinone oxidoreductase subunit I;</td></tr>
<tr><td>sprot_ec</td><td>7.1.1.-;</td></tr>
<tr><td>sport_go</td><td>GO:0005886;GO:0051539;GO:0005506;GO:0050136;GO:0048038;</td></tr>
<tr><td>sport_kos</td><td>K00338;</td></tr>
<tr><td>sport_target</td><td>db:uniprot_sprot|sp|Q1LPV5|NUOI_CUPMC 1 163 evalue:4.1e-65 qcov:100.00 identity:100.00;</td></tr>
<tr><td>tigrfam_go</td><td>GO:0050136;GO:0055114;</td></tr>
<tr><td>tigrfam_desc</td><td>NADH-quinone oxidoreductase, chain I;</td></tr>
<tr><td>tigrfam_id</td><td>NuoI;</td></tr>
<tr><td>tigrfam_mainrole</td><td>Energy metabolism;</td></tr>
<tr><td>tigrfam_sub1role</td><td>Electron transport;</td></tr>
<tr><td>tigrfam_target</td><td>db:TIGRFAMs.hmm|TIGR01971 20 141 evalue:2.1e-48 qcov:73.93 identity:52.46 score:152.8 seqT:153.0 name:TIGR01971;</td></tr>
</table>

Annotations are generated and visualized for the complete metagenome, as well as for each individual MAG. Depending on the research question, users can opt to only annotate a few selected MAGs. Alternatively, they could annotate the entire metagenome first and then use one of MetaErg's utility scripts to extract annotations for each individual MAG. While the annotation of the complete metagenome provides insight into a community's taxonomic composition and metabolic potential, analysis of an individual MAG presents this information for a single organism or population.

Because of the size and density of information in metagenome analysis, exploration of the data presents an overwhelming task that often takes many years to complete (Devlin et al., 2018). To address that challenge, MetaErg produces annotation summary results in various formats. The interactive HTML interface brings all annotation results together in sortable and searchable tables, collapsible trees, and other graphic representations, enabling intuitive navigation of complex data.

With typically massive metagenomic data, similarity-based functional analysis approaches usually suffer from excessive computation time. To address that, DIAMOND is used instead of BLASTP. Diamond is 500 to 20,000 times faster than Blast search tools with a similar degree of sensitivity. To overcome the computational bottleneck and to speed up the functional annotation process, the most time-consuming steps such as database searching in MetaErg are parallelized. Therefore, they run effectively on multicore processors.

Due to the high diversity and large proportion of uncharacterized microbial taxa in most environmental habitats, many microorganisms from environmental samples have no close reference genomes available. While a blast-like tool can quickly identify very similar genes, more distantly related genes can be missed. A profile HMM-based strategy is better at finding more divergent matches and gains sensitivity by incorporating position-specific information into the alignment process and by quantifying variation at each sequence position (Skewes-Cox et al., 2014). MetaErg relies on both Blast and HMM databases (PFAM, TIGRAMs, Metabolic-hmm, casgenes.hmm, and FOAM). FOAM is a manually curated HMM database for identifying functional genes in environmental metagenomes and transcriptomes. Because FOAM was last updated in 2014, we are implementing the addition of UniRef as an alternative, for the next release of MetaErg. Gene annotations such as the EC number and KO number, currently provided by FOAM, could be retrieved from UniRef instead.

SignalP and TMHMM are established signal peptide and transmembrane helix prediction tools. Phobius (Kall et al., 2004) is a combined transmembrane topology and signal peptide predictor. Phobius runs faster on the same dataset than SignalP and TMHMM. However, running Phobius on a 64-bit Linux system requires manually changing its source code before running, due to problems with the included decodeanhmm program. For that reason, we did not select Phobius as a dependency for MetaErg.

Taxonomic classification of genes by similarity searches can be misleading because of the uneven representation of taxa in databases. This can lead to a bias towards highly sampled taxa (Kunin et al., 2008). In addition, with the growing size of the databases, searching all available sequence information becomes computationally challenging. To partially overcome this challenge and improve the classification of uncultured organisms, MetaErg classification databases were built based on GTDB, which provides a more even sampling across *Bacteria* and *Archaea*. Because microbial communities usually also comprise *Eukaryotes* and viruses, we have also added protein sequences of unicellular protozoa, fungi, plants (excluding *Embryophyta*), and viruses. Because MetaErg currently uses Prodigal for gene prediction, it is unable to correctly predict protein sequences of *Eukaryotes*. We are currently working on implementing workflows for better predictions of eukaryotic coding sequences, which will become part of the next version of MetaErg. Likewise, effective identification and analysis of viral contigs is currently still lacking and will become part of the next version.

Although advances in metagenomics have enabled a better understanding of microbial phylogenetic and functional gene compositions in microbiomes, it is also desirable to know which genes are actually expressed. This could be visualized based on transcriptomics or proteomics data (White et al., 2016). Currently, MetaErg enables visualization of expression based on proteomics data only. Visualization of transcriptomics data is planned for a future release.

In conclusion, MetaErg is an easy to use and robust metagenome analysis pipeline. It produces comprehensive analysis reports in various formats. The interactive visualizations help to ease the challenge in interpreting complex analysis results. MetaErg is fully open source and portable, available as a docker image, designed to run on moderately sized computational clusters. Its modular architecture enables addition of new functions. In the future, MetaErg will be expanded by adding new functionality focusing on identification and annotation of eukaryotic and viral MAGs, annotation and discovery of gene clusters encoding production of secondary metabolites, and visualization of transcriptomic data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/?term=prjeb19901.

AUTHOR CONTRIBUTIONS

MetaErg was conceived by XD and MS; XD implemented program with the input from MS; and XD and MS wrote the paper.

FUNDING

We acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Foundation for Innovation (CFI), the Canada First Research Excellence Fund (CFREF), Alberta Innovates, the Government of Alberta, Genome Canada, the University of Calgary, and the International Microbiome Center.

ACKNOWLEDGMENTS

We thank Angela Kouris for proofreading this paper.

REFERENCES

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anantharaman, K., Brown, CT., Hug, LA., Sharon, I., Castelle, CJ., Probst, AJ., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*. 7:13219. doi: 10.1038/ncomms13219.

BBairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28 (1), 45–48. doi: 10.1093/nar/28.1.45

Buchfink, B., Xie, C., and Huson D.H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12 (59), 60. doi: 10.1038/nmeth.3176

Burstein, D., Harrington, L. B., Strutt, S. C., Probst, A. J., Anantharaman, K., Thomas, B. C., et al. (2016). New CRISPR-Cas systems from uncultivated microbes. *Nature* 542 (7640), 237–241. doi: 10.1038/nature21059

Bushnell, B. (2014). *BBMap: a fast, accurate, splice-aware aligner*. (Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory).

Chen, I. M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., et al. (2017). IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleics Acids Res.* 45, D507–D516. doi: 10.1093/nar/gkw929

Devlin, J. C., Battaglia, T., Blaser, M. J., and Ruggles, K. V. (2018). WHAM!: a web-based visualization suite for user-defined analysis of metagenomic shotgun sequencing data. *BMC Genomics* 19, 493. doi: 10.1186/s12864-018-4870-z

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7 (10), e1002195. doi: 10.1371/journal.pcbi.1002195

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2014). Pfam: the protein families database : towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkt1223

Haft, D. H., Selengut, J. D., Richter, A. R., Harkins, D., Basu, M. K., and Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* 41, D387–D395. doi: 10.1093/nar/gks1234

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

Kall, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338, 1027–1036. doi: 10.1016/j.jmb.2004.03.016

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 46, D335–D342, D1. doi: 10.1093/nar/gkx1038

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi: 10.1093/nar/28.1.27

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165. doi: 10.7717/peerj.1165

Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). The MetaCyc Database. *Nucleic Acids Res.* 30 (1), 59–61. doi: 10.1093/nar/30.1.59

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). "MG-RAST, a metagenomics service for analysis of microbial community structure and function," in *Microbial Environmental Genomics (MEG)*. Eds. F. Martin and S. Uroz (New York, NY: Humana Press), 207–233. doi: 10.1007/978-1-4939-3369-3_13

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., et al. (2017). Assessing species biomass contributions in microbial communities *via* metaproteomics. *Nat. Commun.* 8 (1), 1558. doi: 10.1038/s41467-017-01544-x

Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., et al. (2009). The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 37, D216–D223. doi: 10.1093/nar/gkn734

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

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

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi: 10.1093/nar/gkh152

Li, P. E., Lo, C. C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., et al. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleics Acids Res.* 45 (1), 67–80. doi: 10.1093/nar/gkw1027

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI Metagenomics in 2018: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleics Acids Res.* 46, D526–D735. doi: 10.1093/nar/gkx967

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27 (5), 824–834. doi: 10.1101/gr.213959.116

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi: 10.1038/nbt.4229

Prestat, E., David, M. M., Hultman, J., Tas, N., Lamendella, R., Dvornik, J., et al. (2014). FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res.* 42:e145 (19). doi: 10.1093/nar/gku702

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 (Database issue), D590–D596. doi: 10.1093/nar/gks1219

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 14 (15), 2068–2069. doi: 10.1093/bioinformatics/btu153

Skennerton, C. (2016). [27 May 2016]. Minced—mining CRISPRs in environmental datasets. https://github.com/ctSkennerton/minced.

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., and DeRisi, J. L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* 9, e105067. doi: 10.1371/journal.pone.0105067

Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018). DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* 34 (6), 1037–1039. doi: 10.1093/bioinformatics/btx713

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleics Acids Res.* 44 (14), 6614–6624. doi: 10.1093/nar/gkw569

Vallenet, D., Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., et al. (2017). MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleics Acids Res.* 45, D517–D528. doi: 10.1093/nar/gkw1101

Wheeler, T. J., and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

White, R. A. III, Callister, S. J., Moore, R. J., Baker, E. S., and Jansson, J. K. (2016). The past, present and future of microbiome analyses. *Nat. Protoc.* 11, 2049–2053 . doi: 10.1038/nprot.2016.148

Ye, Y., and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* 5, e1000465. doi: 10.1371/journal.pcbi.1000465

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

frontiers
in Genetics

HYPOTHESIS AND THEORY
published: 15 October 2019
doi: 10.3389/fgene.2019.00957

Review, Evaluation, and Directions for Gene-Targeted Assembly for Ecological Analyses of Metagenomes

*Jiarong Guo[1]†, John F. Quensen[1]†, Yanni Sun[2], Qiong Wang[1], C. Titus Brown[3], James R. Cole[1] and James M. Tiedje[1]**

[1] Center for Microbial Ecology, Michigan State University, East Lansing, MI, United States, [2] Department of Electronical Engineering, City University of Hong Kong, Kowloon, Hong Kong, [3] Department of Population Health and Reproduction, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:
Bas E. Dutilh,
Utrecht University, Netherlands

Reviewed by:
Yuzhen Ye,
Indiana University Bloomington,
United States
Arthur Gruber,
University of São Paulo, Brazil

***Correspondence:**
James M. Tiedje
tiedjej@msu.edu

†These authors have contributed equally to this work

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

Received: *23 April 2019*
Accepted: *09 September 2019*
Published: *15 October 2019*

Citation:
Guo J, Quensen JF, Sun Y, Wang Q, Brown CT, Cole JR and Tiedje JM (2019) Review, Evaluation, and Directions for Gene-Targeted Assembly for Ecological Analyses of Metagenomes. Front. Genet. 10:957. doi: 10.3389/fgene.2019.00957

Shotgun metagenomics has greatly advanced our understanding of microbial communities over the last decade. Metagenomic analyses often include assembly and genome binning, computationally daunting tasks especially for big data from complex environments such as soil and sediments. In many studies, however, only a subset of genes and pathways involved in specific functions are of interest; thus, it is not necessary to attempt global assembly. In addition, methods that target genes can be computationally more efficient and produce more accurate assembly by leveraging rich databases, especially for those genes that are of broad interest such as those involved in biogeochemical cycles, biodegradation, and antibiotic resistance or used as phylogenetic markers. Here, we review six gene-targeted assemblers with unique algorithms for extracting and/or assembling targeted genes: Xander, MegaGTA, SAT-Assembler, HMM-GRASPx, GenSeed-HMM, and MEGAN. We tested these tools using two datasets with known genomes, a synthetic community of artificial reads derived from the genomes of 17 bacteria, shotgun sequence data from a mock community with 48 bacteria and 16 archaea genomes, and a large soil shotgun metagenomic dataset. We compared assemblies of a universal single copy gene (*rplB*) and two N cycle genes (*nifH* and *nirK*). We measured their computational efficiency, sensitivity, specificity, and chimera rate and found Xander and MegaGTA, which both use a probabilistic graph structure to model the genes, have the best overall performance with all three datasets, although MEGAN, a reference matching assembler, had better sensitivity with synthetic and mock community members chosen from its reference collection. Also, Xander and MegaGTA are the only tools that include post-assembly scripts tuned for common molecular ecology and diversity analyses. Additionally, we provide a mathematical model for estimating the probability of assembling targeted genes in a metagenome for estimating required sequencing depth.

Keywords: gene-targeted assembly, microbial ecology, gene-centric assembly, Xander, MegaGTA

Abbreviations: pHMM, protein profile hidden Markov model; DBG, *de Bruijn* graph; kmer, subsequence of length k; OTU, operation taxonomic units; Gbp, 1 billion base pairs; GB, 1 billion bytes; *rplB*, the gene encoding 50S ribosomal large subunit L2; *rpsC*, the gene encoding 30S ribosomal small subunit protein S3; *nifH*, the gene encoding nitrogenase reductase; *nirK*, the gene encoding nitrite reductase.

INTRODUCTION

Metagenomics, involving the shotgun sequencing of DNA extracted from environmental samples, has transformed our understanding of microbial ecology in many environments (Qin et al., 2010; Howe et al., 2014; Sunagawa et al., 2015). This method produces reads from random DNA fragments from genomes in the community (National Research Council, 2007). Thus, it has the potential to both overcome the primer bias issue of amplicon-based methods and to provide a broader functional picture of the sampled microbiome (Frank et al., 2008; Klindworth et al., 2013; Guo et al., 2016). To accomplish this, the reads need to be assembled and/or binned in a meaningful way.

Global assembly and local (targeted) assembly are two common strategies for assembling shotgun reads. Global assembly attempts to recover most if not all genomes in metagenomes and has become a common step for shotgun metagenomic analyses. Many assemblers have been developed for this task including MetaVelvet, IDBA-UD, MEGAHIT, and metaSPAdes (Namiki et al., 2012; Peng et al., 2012; Li et al., 2015; Nurk et al., 2017). While major improvements have been made in recent years, global assembly still faces challenges including repeats, sequencing errors, uneven coverage, and the sheer size of data sets, especially for complex environments such as soil (Li et al., 2015; Nurk et al., 2017; Sczyrba et al., 2017). Many studies, however, only focus on genes involved in certain pathways such as the biogeochemical cycles or other genes that are directly responsible for important ecological functions. In these cases, it is not necessary to assemble all of the shotgun metagenomic data, and local assemblers that target these genes of interest may be more advantageous because they focus computational efforts only on assembly of alleles of a specified gene. In parallel with global assembly, significant progress with local assembly has been made in the last 5 years (Zhang et al., 2014; Wang et al., 2015; Alves et al., 2016; Gregor et al., 2016; Zhong et al., 2016; Huson et al., 2017; Li et al., 2017). This has enabled microbial ecologists to recover full-length (or nearly so) marker genes of phylogenetic or functional interest from complex environmental samples without relying on PCR primers that often amplify only partial gene sequences and have well-known biases (Frank et al., 2008; Klindworth et al., 2013; Guo et al., 2016) resulting in more reliable taxonomic assignments and microbial community diversity analyses. Although the target of local assembly can be any genomic segments including genes, gene cassettes, plasmids, or even whole genomes, we focus on protein-coding gene-targeted assemblers in this review.

There are potential problems with all assembly-based methods. First, the assembled contigs may be chimeric. While some of these can be detected and removed using paired-end information, there is no method to verify all *in silico* (Edgar, 2016). Second, sequence variations from closely related strains are collapsed in the assembly process (Awad et al., 2017; Nurk et al., 2017; Brown et al., 2018). Thus, the assembled contigs are not suitable for SNP (single-nucleotide polymorphism), primer design, or diversity analyses that involve fine taxonomic (species or strain) level discrimination. Third, rare members do not have enough coverage to assemble. All of the above are more problematic in complex metagenomes from environments that have high diversity with many closely related strains and many strains with low coverage (Howe et al., 2014).

Gene-targeted assemblers have potential advantages over global assemblers that may minimize such problems: (1) assembly guided by reference can reduce chimera formation and assembly errors arising from sequencing errors; (2) better efficiency from reduced graph and/or search space enables gene-targeted assemblers to use more sophisticated algorithms to explore micro-heterogeneity of closely related strains (Wang et al., 2015; Huson et al., 2017); and (3) the most common current genome binning approach, which relies on the results from global assembly, misses even more low coverage members than targeted assembly since only bins with high completeness and low contamination are usually selected for downstream analyses (Brown et al., 2018). While many gene-targeted assemblers reviewed here demonstrated better performance than global assembly in their original studies (Zhang et al., 2014; Wang et al., 2015; Huson et al., 2017; Li et al., 2017), continual improvements in global as well as gene-targeted assemblers may result in different performances which may also depend on data size, quality, and gene characteristics. Here, we focus on comparing gene-targeted assemblers rather than gene-targeted assemblers *versus* global assemblers.

While assembly outputs are linear sequences, assembly processes require more sophisticated graph data structures. The two most common data structures are *de Bruijn* graph (DBG) and overlap graph (Myers, 2016). The DBG method first chops reads into even smaller kmers and then builds a graph connecting kmers that share k – 1 bases. The overlap graph method first finds overlaps (larger than a length cutoff) among all reads and then connects reads based on the overlapping information (Peltola et al., 1984; Simpson and Durbin, 2012). Earlier methods for constructing the overlap graph required all-against-all pairwise read comparisons and thus were computationally expensive. Recently, efficient overlap detection methods using advanced data structures such as FM-index and Burrows and Wheeler Transform (Lippert et al., 2005; Simpson and Durbin, 2012) have been developed and make overlap detection highly efficient. The DBG is anti-intuitive by breaking down the reads first, but it achieves faster CPU time by avoiding the expensive all-against all pairwise comparisons since the connections among the kmers are implicit (there are only eight possible neighboring kmers for each kmer by extending A, T, C, or G on both ends). DBG is very sensitive to sequencing errors because each sequencing error can cause k spurious kmers and greatly increase the complexity of the graph. Overall, for global metagenomic assembly the overlap graph works well with long reads by preserving the integrity of the reads, whereas the DBG fits well with the massive amounts of short reads that second-generation sequencing platforms produce (Simpson and Pop, 2015; Myers, 2016).

Protein-Coding Gene-Targeted Assemblers

Here, we review and compare the efficiencies and assembly quality of several gene-targeted assembly tools: Xander, MegaGTA, SAT-Assembler, HMM-GRASPx, GenSeed-HMM, and MEGAN's gene-centric assembler (Zhang et al., 2014;

Wang et al., 2015; Alves et al., 2016; Huson et al., 2016; Zhong et al., 2016; Huson et al., 2017; Li et al., 2017). Our goal is to give biologists an easy-to-understand review on the gene-targeted assembly algorithms. This is not a complete list of all gene-targeted assemblers. Rather, our selection criteria were (1) unique innovations in assembly algorithms and (2) scalability with large shotgun metagenomic data.

The tools reviewed here use a wide range of algorithms and can be divided into two main categories (**Table S1**): (1) read filtering, potentially iteratively, using sequences or pHMMs as search queries, and (2) assembly by alignment, where pHMMs are used for guiding graph traversal in assembly. Among the tools reviewed, pHMM-GRASPx, GenSeed-HMM, and SAT-Assembler belong to first category. HMM-GRASPx and GenSeed-HMM use iterative read-filtering steps to potentially elongate nascent contigs and then apply third party tools for assembly, while SAT-Assembler has a novel assembly algorithm. MEGAN's gene-centric assembly function is similar to the first category except that it first aligns all reads against NCBI-nr and subsets reads that align to target genes. Further, Xander and MegaGTA belong to the second category and share a novel pHMM-guided graph traversal algorithm.

1) Xander

Xander combines a DBG with a protein profile Hidden Markov Model (pHMM) built from a reference set of target gene sequences. The probabilities from the pHMM guide gene assembly (Wang et al., 2015). The DBG is encoded as a lossy (approximate) data structure which compresses the sequence data (Pell et al., 2012). The memory needed for this data structure is dependent on the data complexity, not total data size. Xander requires the user to specify the amount of memory before compression. If too little memory is specified for an accurate compression, the user will need to re-run the time-consuming compression. Xander searches start at all nucleotide kmers with sequences that potentially encode short protein sequences found in one or more target gene reference sequences. These starting kmers are extended in both 5' and 3' directions using the encoded pHMM probabilities to find high-probability paths in the graph structure, analogous to the way a pHMM is used to find high-probability alignments in a (linear) DNA sequence. The traversal advances three graph nodes (three kmers) at a time (one codon) to select a single reading frame for the pHMM. Xander uses the "A*" algorithm (Hart et al., 1968) to find the path with the highest probability and can also find multiple paths from one start, which is important when studying allelic diversity, using the modified Yen's K shortest path algorithm (Yen, 1971; Lawler, 1972), which is further modified to require each additional path to contain at least one unique kmer. Therefore, pHMM-guided graph traversal not only reduces the search space compared to global assembly but also provides a probability measure analogous to the familiar BLAST bits score for how likely a contig would have matched the pHMM by chance and thus reduces assembly error.

To assemble sequences, Xander requires forward and reverse pHMMs built from a relatively small set of protein sequences (e.g., 117 for *rplB*) that capture the diversity of the target gene, and a larger set of aligned protein sequences (1,743 for *rplB* but can be several thousands) for finding starting kmers. The current Xander package includes models for the single copy ribosomal protein gene *rplB* and a few N cycle genes (AOA, AOB, *nifH*, *nirK*, *nirS*, *norB_cNor*, *norB_qNor*, *nosZ_cladeI*, and *nosZ_cladeII*). A tutorial is provided for preparing the required pHMMs and references for additional genes (https://github.com/rdpstaff/Xander_assembler#per-gene-preparation-requires-biological-insight).

Another unique aspect of Xander is that it is designed for microbial diversity analyses and thus includes post-assembly utilities such as chimera checking, *de novo* OTU clustering, taxonomic classification (the nearest neighbor in the reference database with percent identity), and quantification. After assembly, the contigs are clustered at 99% to remove redundancy, and the chimeras are removed by UCHIME (Edgar et al., 2011). For these post-assembly tasks, Xander requires a large set of protein sequences with taxonomy information in the descriptions (usually the same as those used for finding starting kmers) and a comparable set of nucleotide sequences.

2) MegaGTA

MegaGTA is designed based on Xander's analysis framework and claims several improvements: (1) MegaGTA uses a different space-efficient variant of DBG, the succinct *de Bruijn* graph (sDBG) that was first implemented in the popular global assembly tool MEGAHIT (Li et al., 2015). The sDBG is highly parallelizable and can also be used to build an iterative DBG (Peng et al., 2010), which is difficult to achieve with the bloom filter employed by Xander. The iterative DBG allows the use of multiple kmer sizes, increasing sensitivity and specificity. (2) Xander is designed to remove erroneous kmers caused by sequencing errors by filtering out kmers with low abundance but then keeps single-copy "mercy-kmer" (Li et al., 2015) if they are the only kmers connecting two abundant kmers in a read for the purpose of retaining low abundance species in metagenomes. These are common in complex environments, but this could potentially reintroduce kmers that are sequencing errors. Although pHMM-guided graph traversal should reduce the chance of erroneous kmers entering assemblies, MegaGTA does penalize kmers with low coverage in the guided assembly step. This reduces assembly error from sequencing errors but might also introduce bias against low abundant members. Overall, MegaGTA achieves better sensitivity and specificity, although its memory requirement can still be a hindrance for large and complex metagenomes.

3) SAT-Assembler

Similar to Xander and MegaGTA, SAT-Assembler also uses pHMM, but it is a string graph–based assembler that includes two main steps. The first step searches for target gene fragments in reads using pHMM with HMMER3 with a permissive cutoff (e-value cutoff of 1,000), which greatly reduces the input data size for the next step without losing sensitivity. The second step builds a string graph for each targeted gene and assembles contigs. The read alignment location information against the model from the first step is used to guide the overlap calculation among reads. Multiple types of information such as paired ends, overlap connection, and coverage are used to guide graph traversal and avoid chimeras. Contigs are merged into scaffolds using paired-end information as the final step. To run SAT-Assembler, a file containing pHMMs of targeted genes is required. The pHMM for a specific gene can be built from aligned protein sequences of the gene using the hmmbuild command in HMMER3 (Eddy, 2009). Additionally, SAT-Assembler is also designed to work with pHMMs in the Pfam database, which has ~ 18,000 pHMMs in version 32.0 and covers ~ 80% of protein sequences in UniProtKB (Finn et al., 2016; Schaeffer et al., 2017).

As mentioned briefly above, Xander/MegaGTA and SAT-Assembler use pHMMs in very different ways. In Xander/MegaGTA, pHMMs are used to guide graph's traversal in DBG. Although the graph traversal space is reduced to those paths related to the target gene, it is still computationally expensive (CPU time and memory) to load all reads into the graph and identify all starting kmers in a large graph. In contrast, SAT-Assembler uses pHMMs to filter reads belonging to target genes as a data reduction step and then uses the reduced dataset to build the assembly graph, thus greatly reducing the memory and CPU cost of graph building. SAT-Assembler further uses read pHMM alignment information to speed up overlap computation among reads for building string graphs. It, however, does not apply pHMM to guide graph traversal on the resulting string graph, which could potentially improve the assembly.

4) HMM-GRASPx

HMM-GRASPx is also pHMM-based, but it integrates many tools including gene callers (MetaGeneAnnotator/FragGeneScan) (Noguchi et al., 2008; Rho et al., 2010), HMMER3 (Eddy, 2009), nucleotide sequence assembler (SPAdes) (Nurk et al., 2017), and protein sequence assembler (SFA-SPA) (Yang et al., 2015). Its core algorithm, iterative search and assembly, is based on an overlap graph in protein space and hence can increase the sensitivity of gene identification. Short reads are not ideal for gene identification because they may not have enough information to be recognized as the target gene. HMM-GRASPx tackles this problem by iterative search and assembly. Intuitively, homologous protein sequences translated from reads with low sequence identity could be identified by being assembled first with other high identity reads into longer contigs. More specifically, (1) overlaps among reads are firstly computed, (2) reads with high pHMM alignment scores are identified and used as starting contigs, (3) contigs are extended using overlapping reads, and (4) the extended contigs are aligned with pHMM to decide whether to continue extending. If the alignment score is below a certain threshold or there are no more overlapping reads, then the extension stops; (5) the resulting contigs are assembled again based on their overlap; and (6) finally, reads from the target gene are retrieved by mapping them to the assembled gene contigs. This core algorithm functions both as a finder and assembler. HMM-GRASPx's authors suggest that, for quantitative results, the identified contigs be assembled with another program, i.e., SPAdes for nucleotide and SFA-SPA for protein reads. This is because the algorithm outputs all possible contigs to increase sensitivity and thus can produce redundant assemblies. However, it should be possible to simply remove the redundant contigs, which would improve the overall computational efficiency.

5) GenSeed-HMM

GenSeed-HMM applies an iterative assembly and extension strategy similar to that used by HMM-GRASPx. The key difference is GenSeed-HMM can extend beyond the gene boundaries, while HMM-GRASPx will automatically stop extending when the pHMM alignment score drops. GenSeed-HMM has the advantage of being able to use nucleotide, protein sequences, or pHMMs as references, which gives the users more flexibility. Internally, it applies BLASTn with nucleotide references and TBLASTN for protein references to search against the (nucleotide) reads, and hmmsearch for pHMM search of the translated reads. At the assembly step, it uses third party assembly tools such as SOAPdenovo, ABySS, and CAP3 (Huang and Madan, 1999; Simpson et al., 2009; Li et al., 2010; Luo et al., 2012), and the choice of third party assembly tools might have an impact on its overall computational efficiency and assembly quality. For contig extension iterations, contig ends are extracted and used as new references for the next search iteration. If no contigs are extended, it will trim the extended part from the previous iteration and try new extension up to three iterations. Once a contig reaches or exceeds the maximum length set, it will not be included in subsequent iterations. GenSeed-HMM is not a typical gene-targeted assembler since its contigs may extend beyond gene boundaries. This makes it useful to study the nearby genes (genomic context) of the target gene. For marker gene–based microbial diversity studies, however, the parts beyond the gene boundaries would have to be trimmed before further analyses.

6) MEGAN-Assembler

MEGAN assembler is part of MEGAN version 6 (Huson et al., 2016; Huson et al., 2017), and its key algorithm is protein alignment-guided assembly, an overlap graph–based method. It requires an all against all pairwise alignment of query metagenomes and reference database such as NCBI-nr using BLAST or DIAMOND (Altschul et al., 1997; Buchfink et al., 2015) as the first step, the same as all other analyses in MEGAN. MEGAN utilizes the above alignment information to find the overlap among reads based on their alignment to the same target references and further constructs overlap graphs based on 100% sequence match in the overlapped portion of the alignment. In this way, MEGAN avoids the expensive computation of all against all comparisons among query reads for constructing overlap graphs (similar to SAT-Assembler). Further, MEGAN weights overlap graph edges (connection between reads) by overlap sizes and then traverses the graph by finding an acyclic path with a maximum weight. It reports contigs with a minimal length, removes the reads used for the assembled contigs in overlap graphs, and iterates the above process until no more paths remain. Contigs are further extended if two contigs have overlap and an overlap identity larger than a certain thresholds (by default 20 bp and 98%, respectively). Although inducing the read overlap from alignment against references is a good strategy to improve computational efficiency, the first step of all *vs*. all comparison of query to NCBI-nr is still a daunting task for large metagenomes.

METHODS

Data

We evaluated the performance of these gene-targeted assemblers using three data sets. The synthetic data consisted of 150-bp single reads without errors generated from the 17 genomes in **Table S2** using Grinder (Angly et al., 2012) with the parameters "-rd 150 -cf 10" to give 10X coverage of each genome. The seven species of *Pseudomonas* were selected as a challenge for assemblers regarding their production of chimeric contigs. The mock community data, generated from a mixture of known amounts of gDNA from 16 archaeal and 48 bacterial strains (Shakya et al., 2013), consisted of 100-bp paired Illumina reads downloaded from NCBI as run SRR606249. These reads were trimmed using fastq-mcf (version 1.04.662) (http://code.google.com/p/ea-utils) with the parameters "-q 30 -l 50 -w 4 -x 10 -max-ns 0 -X." The soil metagenome sample was sample C1 that was included in the original Xander paper (Wang et al., 2015) and is available from NCBI as run SRR3989263. Fifty million reads sampled from C1 were trimmed with fastq-mcf with the same parameters above and converted to FASTA format to give 33.7 million paired reads designated C1-50M.

Programs

Xander is included in RDPTools, which is available as source on GitHub (https://github.com/rdpstaff/RDPTools). It requires Python 2.7+, Java 1.6+, HMMER 3.1 (http://hmmer.janelia.org), and UCHIME (http://drive5.com/usearch/manual/uchime_algo.html). All of these dependencies may be met by instead installing the Bioconda package from https://bioconda.github.io/recipes/rdptools/README.html. Instructions for Xander are available at https://github.com/rdpstaff/Xander_assembler and https://john-quensen.com/workshops/workshop-2/xander. We installed RDPTools from source. All required reference files for *rplB*, *nifH*, and *nirK* are included in the installation.

Two of Xander's parameters depend on the input file size. We set FILTER_SIZE to 32, 36, and 38, and MAX_JVM_HEAP to 4G, 12G, and 64G for the synthetic, mock, and C1-50M data, respectively. We set MIN-COUNT to 1 and left all other parameters at their default values for all cases. Resulting false-positive error rates were always less than 3.20E−05.

MegaGTA is a re-write in C++ of the first two portions of Xander: build and find. It may be installed from source from https://github.com/HKU-BAL/megagta or as a Bioconda package from https://bioconda.github.io/recipes/megagta/README.html. MegaGTA requires RDPTools. If installed from source, RDPTools is included. If installed from Bioconda, RDPTools must be installed separately. We installed the Bioconda package.

We limited the available memory for MegaGTA to 19.2G for the synthetic data and left all other parameters at their default values, including memory, for the other data sets. Memory is set as a fraction (0.8 by default) of available memory. The gene_list.txt configuration file used pointed to the for_enone.hmm, rev_enone.hmm, and ref_aligned.fasta files for each gene (*rplB*, *nifH*, and *nirK*) in the RDPTools/Xander_assembler/gene_resource directory.

We installed SAT-Assembler from the forked version on GitHub at https://github.com/jiarong/SAT-Assembler, following the instructions on that web page. Older versions of SAT-Assembler on SorceForge.net and at https://github.com/zhangy72/SAT-Assembler no longer work because of updates to some of the modules the program requires. For this program, HMM-GRASPx and GenSeed-HMM, we used pHMMs downloaded from the FunGene web page (http://fungene.cme.msu.edu/).

We installed HMM-GRASPx from https://sourceforge.net/projects/hmm-graspx/ and followed the directions under the Files tab on that page. To generate input files for HMM-GRASPx, we ran FragGeneScan with parameters "-complete 0 -train illumine_5 –thread 4." For HMM-GRASPx, we left all parameters at their default values.

We installed the Linux version of MEGAN and its auxiliary mapping files from http://ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html. Use of MEGAN for gene-centric assembly from metagenomic data requires that all sequences are first aligned against NCBI's non-redundant protein database (NCBI-nr). We used DIAMOND (Buchfink et al., 2015) (https://github.com/bbuchfink/diamond) because of its speed and output format 100 since the resulting daa (DIAMOND alignment archive) files are more rapidly imported into MEGAN. We "meganized" the data files using the protein accession to InterPro mapping file acc2interpro-June2018X.bin downloaded from the MEGAN site and the command line tool daa-meganizer. For both DIAMOND and MEGAN assembler, we used the default values for all parameters.

GenSeed-HMM is a Perl script available at https://sourceforge.net/projects/genseedhmm/. It operates by making calls to a variety of third-party tools including BLAST+, hmmsearch, EMBOSS, bowtie, and at least one assembler. We used the ABySS assembler for all of our tests with this program. We used Conda to create an environment containing these programs and their dependencies and ran GenSeed-HMM from within this environment. An YML file for creating the same environment is available at https://github.com/jfq3/Virtual-Environments.

Assembly Quality

We evaluated two aspects of assembly quality: (1) contigs should capture all target gene sequences known to be in the data (sensitivity), and (2) contigs should not include irrelevant sequences (specificity). Both aspects were evaluated by conducting a BLAST search of contigs against target gene sequences extracted from the genomes, or in the case of the soil sample C1-50M against NCBI-nr. Sequence similarity was defined as "alignment length" * identity/"length of shorter sequence." Some contigs were too different from the target sequences to appear in the BLAST results. The relationships of such contigs to the target genes were investigated by searching against NCBI-nr and/or against the genomes themselves and viewing the alignment in NCBI's genome browser. Potentially chimeric sequences assembled from the synthetic and mock data were also flagged by UCHIME using target gene sequences extracted from the genomes as the reference.

To make these tests comparable among assemblers, we compared comparable contigs. For Xander and MegaGTA, we used the intermediate file “_prot_merged_rmdup.fasta.” Post-assembly *per se*, Xander and MegaGTA outputs are normally processed through a pipeline that removes potential chimeras and short sequences and clusters the remaining sequences at a user-defined distance, thus decreasing sequence variation in their final outputs. The file “_prot_merged_rmdup.fasta” has not been subjected to these processes and contains all unique contigs assembled. To investigate chimeras produced by Xander and MegaGTA, corresponding nucleotide sequences were selected from the “nucl_merged.fasta” files; these files are all nucleotide contigs assembled. As well as testing SAT-Assembler and GenSeed-HMM output directly, we also removed duplicate sequences and filtered to a minimum length of 450 bp (using RDPTool’s rm-dupseq command) to produce results more comparable to Xander’s and MegaGTA’s “_prot_merged_rmdup.fasta” files. We also compared MEGAN results filtered to the same minimum length.

Sequencing Depth Estimation

In genome sequencing, the relation between sequencing depth and genome coverage is already a well-studied problem. Lander–Waterman statistics (Lander and Waterman, 1988) show that with “L” as read length, “N” as number of reads, and “G” as genome length (much larger than read length), the average coverage of genome (“a”) is “LN/G,” and the probability of each base not being covered (“p”) is “e^{-a}.” In the context of metagenomics, however, a targeted species is only “R” (relative abundance) of the total community, so “a” (the average coverage of genome) should be redefined as “LNR/G” (we assume that all species have the same genome size, “G,” to simplify the problem). We can further deduce that the probability (“P”) of at least “M” continuous positions (a contig with at least “M” bp) in a target gene with a size of “S” bp being covered is:

$$P=\sum_{i=M}^{S}(S-i+1)p^{S-i}(1-p)^{i}$$

Further, the above only considers whether a position is covered but not the read overlaps that are needed for assembly. In DBG graph with kmer size of “k,” the minimal overlap required for two reads to connect is “$k-1$.” To account for the “$k-1$” overlap in either DBG or overlap graph, we can simply define the effective read to be the first “$L-(k-1)$” position of each read, so when one shortened read follows right after where a preceding one ended, they effectively have an overlap of “$k-1$.” Therefore, “p” can be redefined as the probability of a position not being covered by reads of effective length (“$L-k+1$”) with the value:

$$p=e^{-(L-k+1)NR/G}$$

To evaluate the effect of sequencing depth on gene-targeted assembly, we first evenly divided our soil metagenome (C1) into 2, 4, 8, 16, and 32 subsamples. For each sample, we ran Xander to assemble *rplB* with the same parameters mentioned above. The coverage information was retrieved from mean kmer coverage in “_rplB_45_coverage.txt” output file. We also included *rpsC* as a confirmation of *rplB* results. The reference files of *rpsC* for Xander can be downloaded from http://doi.org/10.5281/zenodo.1410823 (Guo, 2018).

RESULTS

Time and Memory Requirements

Comparisons of computer time and memory resources required are complicated by the programs having different prerequisites and end points. Overall, SAT-Assembler was the most efficient requiring less than 6-min wall time and only 78 MB of memory to process the synthetic data for *rplB* (**Table 1**). SAT-Assembler stops short of providing quantitative results allowing sample comparisons as Xander does; such further processing would be close to that for MegaGTA’s post-processing step. Xander’s three steps took only slightly longer (7 min 31s) to provide quantitative results but required approximately 1.5 GB of memory. Xander’s build step is considered a bottleneck because it is not multithreaded, and MegaGTA is advertised as advancement over Xander in part because of greater speed. This is true only for wall time and if enough threads are used; the actual CPU time (78 min) was much greater than Xander’s but did require slightly

TABLE 1 | Time and memory requirements for processing the synthetic data for *rplB*. Except for MEGAN BLAST/DIAMOND performed on MSU’s cluster, all times are for running on an HP ProBook 450 G5 with Intel i7-8550U CPU and 32 Gb RAM running Ubuntu 18.04 LTS.

Program	Stage	Threads	Wall timehh:mm:ss	CPU timehh:mm:ss	Peak memory (KB)
Xander	Build	1	00:03:52	00:03:57	736,860
	Find	4	00:00:57	00:04:48	1,512,728
	Search	4	00:02:42	00:04:28	867,776
MegaGTA	Main	8	00:10:06	01:15:02	1,133,248
	Post-processing	4	00:00:47	00:02:16	729,624
FragGeneScan		4	00:24:20	01:29:15	65,356
HMM-GRASPx		4	00:05:28	00:05:28	8,159,504
SAT-Assembler		NA	00:05:55	00:06:38	77,620
MEGAN	Diamond	8	14:38:57	95:11:48	19,810,188
	Meganize	NA	00:05:46	00:15:57	21,659,968
	Assembly	NA	00:00:03	NA	NA
GenSeed-HMM		4	00:07:46	00:16:57	1,425,368

less memory. The memory requirement for GenSeed-HMM was comparable to that of Xander, but the processing time was approximately twice as long without including any of the post-processing steps required for making sample comparisons.

The pre-processing required by HMM-GRASPx and MEGAN made them much less efficient to implement. HMM-GRASPx requires that all fragments first be translated into peptide reads by FragGeneScan or MetaGeneAnnotator. Furthermore, to obtain accurate quantitative results, the authors recommend that the contigs be re-assembled by another program; time and memory requirements for that process are not included in **Table 1**. MEGAN is by far the least efficient, requiring that all fragments first be aligned against NCBI's non-redundant protein database. For this task, DIAMOND is preferred over BLAST due to its much greater speed (still required over 95-h CPU time), but the speed comes with a higher memory requirement (20 GB).

Assembly Quality Tested With Synthetic Data

GenSeed-HMM was the most successful at capturing exact matches to the *rplB* genes in the synthetic data, matching all 17 with 100% identity (**Table 2**). HMM-GRASPx, MEGAN, and SAT-Assembler did nearly as well, matching 16 of the sequences at 100% identity. HMM-GRASPx missed *Pseudomonas putida* while MEGAN missed *Lacunisphaera limnophila* even at a lower 97% identity threshold. Many of the exact matches produced by HMM-GRASPx, MEGAN, and GenSeed-HMM were short; however, they captured only about half of the target genes if comparisons were restricted to contigs of at least 450 nucleotides. Xander and MegaGTA were the worst at producing exact matches, capturing only 12 of the 17 genes at 100% identity.

These same two assemblers were the best, however, at excluding irrelevant sequences; all 28 contigs were at least 96% identical to *rplB* gene sequences, and all 17 taxa were captured at a 97% identity threshold. HMM-GRASPx also did well, with only 5% of its assemblies having BLAST matches to *rplB* of less than 97% identity. MEGAN, on the other hand, assembled 32 contigs (58% of the total) that were perfect matches to portions of the reference genomes but entirely unrelated to *rplB*, and 58 to 60% of the SAT-Assembler assemblies had less than 97% identity to *rplB* genes in the synthetic data. GenSeed-HMM also assembled some sequences unrelated to the target sequences.

Except for SAT-Assembler, all tools assembled contigs matching all six *nifH* (nitrogenase reductase) sequences present in the synthetic data with at least 97% identity (**Table S3**). SAT-Assembler did not match any of the reads to *nifH* and so did not assemble any contigs for the gene. MEGAN and GenSeed-HMM also produced high proportions of contigs (11 of 20 and 71 of 127, respectively) unrelated to nifH sequences in the synthetic data.

HMM-GRASPx, MEGAN, SAT-Assembler, and GenSeed-HMM all assembled contigs with 100% identity to all four *nirK* (nitrite reductase) sequences present in the synthetic data (**Table S4**). Xander and MegaGTA performed identically, each producing contigs which matched only two of the *nirK* sequences present in the synthetic data, but with 100% identity. MEGAN, SAT-Assembler, and GenSeed-HMM again produced non-relevant contigs.

Assembly Quality Tested With Mock Data

Overall, MegaGTA was the most successful at assembling *rplB* contigs from the mock data, producing 86 unique contigs of more than 450 bp with at least 97% identity to 46 of the 48 bacterial *rplB* sequences present (**Table 3**). While SAT-Assembler using an overlap length of 40 produced more (1,318) contigs with 100% identities to 47 of the 48 rplB sequences present, most of the contigs were very short. There were only 61 unique contigs of at least 450 bp, and only 13 of these matched expected *rplB* sequences with 100% identity. Xander did nearly as well as MegaGTA, while for MEGAN's contigs, over 450 bp matched only 33 of the *rplB* sequences with at least 97% identity and GenSeed-HMM's matched 28 with 100% identity. All the assemblers produced "missing" contigs, i.e., ones that did not appear in the BLAST tables due to very low sequence similarity to reference sequences. By BLAST to NCBI-nr, all of these produced by Xander, MegaGTA, and SAT-Assembler matched known *rplB* sequences at more than 99% identity. Only one, however, of the 45 produced by MEGAN was related to *rplB*.

TABLE 2 | BLAST summary for *rplB* assembled from the synthetic data. There were 17 *rplB* sequences in the synthetic data. Entries in the % ID columns give the number of taxa matched over the number of contigs that match *rplB* by BLAST identity at the specified percentage.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	28	807–828	0	1	17/27	15/23	12/16	12/12
MegaGTA	28	807–828	0	1	17/27	15/23	12/16	12/12
HMM-GRASPx	63	102–261	0	3	16/60	16/60	16/59	16/59
HMM-GRASPx	0	> =450	–	–	–	–	–	–
MEGAN[1]	55	204–3,822	32	0	16/23	16/23	16/23	16/23
MEGAN[2]	20	453–3,822	11	0	9/9	9/9	9/9	9/9
SAT-Assembler[3]	176	150–997	49	60	17/67	17/50	16/28	16/23
SAT-Assembler[4]	106	465–997	0	58	16/48	15/33	13/14	11/11
GenSeed-HMM[5]	97	32–1,340	4	0	17/93	17/93	17/93	17/93
GenSeed-HMM[6]	9	724–1,340	1	0	8/8	8/8	8/8	8/8

MEGAN[1]: all contigs assembled. MEGAN[2]: contigs filtered to a minimum length of 450 bp. SAT-Assembler[3]: all contigs assembled with an overlap length of 40 bp. SAT-Assembler[4]: contigs were de-replicated, duplicates removed, and filtered to a minimum length of 450 bp. GenSeed-HMM[5]: all contigs assembled; GenSeed-HMM[6]: contigs were filtered to a minimum length of 450 bp.

TABLE 3 | BLAST summary for *rplB* contigs assembled from the mock data. There were 48 bacterial *rplB* sequences in the mock data set. Entries in the % ID columns give the number of taxa matched over the number of contigs that match *rplB* by BLAST identity at the specified percentage.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	95	459–849	2	5	44/88	43/85	40/80	30/30
MegaGTA	94	453–849	2	6	46/86	44/83	42/80	32/32
MEGAN[1]	93	201–1,611	45	1	39/47	39/47	38/46	35/39
MEGAN[2]	50	450–1,611	16	1	33/33	33/33	32/32	28/28
SAT-Assembler[3]	2,765	50–750	751	107	48/1,907	48/1,865	48/1,689	47/1,318
SAT-Assembler[4]	61	458–750	1	18	29/42	27/37	25/31	13/13
GenSeed-HMM[5]	408	31–1,360	60	7/9	47/339	47/330	46/187	43/183
GenSeed-HMM[6]	44	450–1,360	11	1/1	28/32	28/32	27/31	23/27

[1] Data for all MEGAN contigs assembled from reads mapping to IPR005880 using default parameters. [2] Data for MEGAN contigs filtered to a minimum length of 450 bp. [3] All SAT-Assembler rplB contigs assembled from the mock data with an overlap length of 40 bp. Notice that the minimum length is one-half of the read length. [4] SAT-Assembler contigs were assembled with an overlap length of 40 bp, de-replicated, duplicates removed, and filtered to a minimum length of 450 bp. HMM-GRASPx failed to complete with this data set. GenSeed-HMM[5]: all contigs assembled; GenSeed-HMM[6]: contigs were filtered to a minimum length of 450 bp.

GenSeed-HMM and MEGAN did slightly better than Xander and MegaGTA in capturing *nifH* sequences in the mock data (**Table S5**), but both again produced high proportions of unrelated contigs and many of GenSeed-HMM's were very short. As with the synthetic data, SAT-Assembler did not match any of the reads to *nifH* and so did not assemble any contigs for the gene.

SAT-Assembler did assemble *nirK* contigs, matching all five sequences present in the data at 100% identity (**Table S6**), but again, most contigs were short. Only two were over 450 bp, and these matched only one of the five *nirK* sequences in the mock data. GenSeed-HMM did better, producing contigs matching all five target genes with 100% identity even after they were filtered for length, but also a high proportion of contigs unrelated to the *nirK* sequences in the data. MEGAN contigs matched four of the five at 100% identity but also produced a high proportion of unrelated sequences. MegaGTA and Xander produced three and two contigs, respectively, matching two of the target sequences.

Assembly Quality Tested With Soil Metagenome

For the C1-50M shotgun data, GenSeed-HMM produced the most contigs and matched the highest number of *rplB* sequences in NCBI-nr (**Table 4**). But most of the contigs were very short such that over 70% did not match *rplB* with an e-value of less than 10. Only two were over 450 bp. Considering only contigs over 450 bp, MegaGTA produced the most (316), all of which matched *rplB* sequences in NCBI-nr, and Xander was a close second. MEGAN produced far fewer contigs (30), only 3 of which were over 450 bp, and 11 of which were not *rplB*.

Chimera

The synthetic data set was meant to be challenging with regard to chimera formation, especially for *rplB*. Xander, MegaGTA, and SAT-Assembler all produced high proportions of *rplB* chimeras from this data set (**Table S7**). For the first two, chimeras were almost exclusively (10 of 11, over 90%) between species of *Pseudomonas*. For SAT-Assembler, however, approximately one fourth of the chimeras were between different genera, and the proportion of chimeras increased with contig length. None of MEGAN's or GenSeed-HMM's contigs were flagged as chimeras.

The same trend held for the mock data (**Table S8**). Xander and MegaGTA produced fewer *rplB* chimeras than SAT-Assembler, and when they occurred, they were exclusively between species of the same genus. In contrast, approximately 30 to 40% of the chimeras (depending on length) produced by SAT-Assembler were between different genera. As with the synthetic data, none of MEGAN's *rplB* contigs were flagged as chimeras, and only 1 of 408 produced by GenSeed-HMM was a chimera.

Xander and MegaGTA also produced a high percentage of *nifH* chimeras from the synthetic data (**Table S9**), but exclusively

TABLE 4 | BLAST summary for bacterial *rplB* contigs assembled from C1-50M aligned against NCBI-nr. Entries in the % ID columns give the number of taxa matched over the number of contigs that match *rplB* by BLAST identity at the specified percentage.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	269	453–825	0	56/250	11/19	8/16	4/8	3/3
MegaGTA	316	450–825	0	82/290	13/26	12/19	8/11	4/4
MEGAN[1]	30	207–705	11	2/2	14/17	11/14	9/12	9/12
MEGAN[2]	3	462–705	1	2/2	2/2	2/2	2/2	2/2
SAT-Assembler[3]	705	51–436	9	125/207	179/469	154/381	132/316	131/312
SAT-Assembler[4]	0	–	–	–	–	–	–	–
GenSeed-HMM[5]	4340	31–1,058	3109	334/596	311/635	284/562	277/535	273/535
GenSeed-HMM[6]	4	458–1,058	0	2/2	2/2	2/2	1/1	1/1

MEGAN[1]: all contigs assembled. MEGAN[2]: contigs filtered to a minimum length of 450 bp. SAT-Assembler[3]: contigs assembled with an overlap length of 40 bp and de-replicated. SAT-Assembler[4]: contigs assembled with an overlap length of 40 bp were de-replicated and filtered to a minimum length of 450 bp. GenSeed-HMM[5]: all contigs assembled; GenSeed-HMM[6]: contigs were filtered to a minimum length of 450 bp.

between sequences from the same genus. In fact, for Xander, two of the five and for MegaGTA three of the five chimeras were between *nifH* copies within the same species. There were only two other instances of chimera formation. Xander formed a *nifH* chimera with the mock data between strains S2 and C5 of *Methanococcus maripaludis*, and MEGAN formed a *nifH* chimera between *Azotobacter vinelandii* and *A. chroococcum*. There were no *nirK* chimeras from either data set by any of the assemblers.

Sequencing Depth

With the derived model, we estimated that ~ 40 Gbp of sequences is needed to assemble a contig (>450 bp) from a gene with a length of 800 bp in a species that is 0.1% of the metagenome (assuming all genome sizes are 5 Mbp) (**Figure 1**). Additionally, when we evaluated the effect of sequencing depth on assembly by subsampling, we found the number of genes assembled decreased much faster than sequencing depth for both *rplB* and *rpsC* (**Figure 2**).

DISCUSSION

Computer time and memory requirements can be limiting factors in deciding a method to process metagenomic data. SAT-Assembler required the least time and memory because it first selects a limited number of reads related to the target gene to assemble. HMM-GRASPx employs a similar strategy to reduce time and memory requirements, but by relying on FragGeneScan as a pre-step, it requires far more total time. Furthermore, its pHMM alignment at each contig extension is also computationally expensive and slows down the simultaneous search and assembly step. Similarly, GenSeed-HMM bogs down trying to extend both ends of the numerous sequences it finds in a first pass through complex data, and MEGAN's reliance on conducting a BLAST search of all sequences against NCBI-nr makes it computationally very expensive to implement. We were only able to compare assembler performance with an environmental sample by reducing the C1 sample to 50 million reads. The full sample is five times as large, and neither GenSeed-HMM nor DIAMOND BLASTX finished processing the full C1 sample within the 7-day limit on our cluster. By contrast, Xander processing of the full C1 data set, including all post-assembly processing, for all three genes considered here took only 18 h 13 min of wall time (40 h 30 min of CPU time).

SAT-Assembler's savings in resource cost comes at great expense in performance, notably in the production of mostly short contigs. The similarity search step may have missed remote homologs of the references in pHMM despite the loose cutoff used in hmmsearch. Thus, by selecting relatively few reads to assemble, there are not enough left to fill gaps in the gene sequence, i.e., to join the shorter contigs. The same problem is seen with HMM-GRASPx. Since it utilizes all reads (in protein space) in its simultaneous search and assembly algorithm, short contigs might be caused by different factors in its pipeline such as the re-calibration step where locally extended contigs are merged. Xander, MegaGTA, and MEGAN, on the other hand, are able to assemble longer contigs because they work from all reads in the sample (at the cost of much larger memory usage and CPU time to load all data) and might also have more robust algorithms to maximize contig lengths.

Sensitivity is also of paramount importance. Considering the number of target genes matched with 100% identity, GenSeed-HMM scored highest, matching all target sequences in the synthetic and mock data. SAT-Assembler scored nearly as well, not considering *nifH*. It matched all *nirK* genes in both the synthetic and mock data, all *rplB* genes in the synthetic data, and all but one of the 48 bacterial *rplB* genes in the mock data. HMM-GRASPx did as well for the synthetic data and additionally assembled contigs that matched all *nifH* genes in the synthetic data, which is

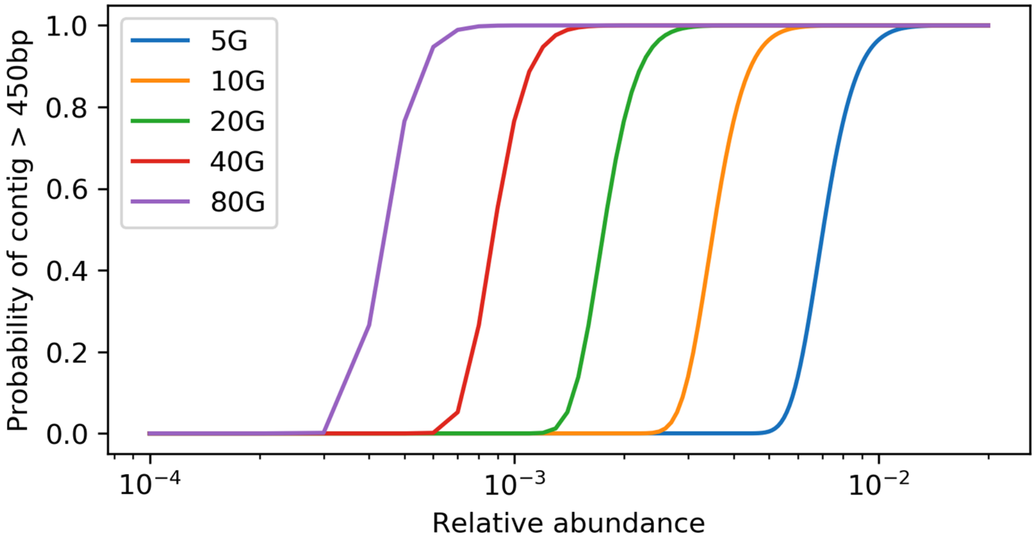


FIGURE 1 | Relation between the probability of having a target gene from a species assembled and the relative abundance of the species at different sequencing depth. X axis is at log10 scale, the target gene length is set to 800 bp, and the minimum contig length is set to 550 bp.

something SAT-Assembler failed to do. MEGAN did just as well with the synthetic data but matched only 35 *rplB* genes in the mock data and only four of the five *nirK* genes in the mock data. It did the best at matching 16 of the 18 *nifH* genes in the mock data at 100% identity. It is easy to understand MEGAN's performance at providing 100% matches to the target genes. Because of the way it works, the contigs it produces are essentially genes in NCBI-nr. As long as a gene in NCBI-nr is well represented in the sample, it is what you get back as the contig. This also means that MEGAN is less likely to capture novel gene diversity in environmental samples. Thus, with different datasets, different genes, and identity cutoff, it is difficult to find the tool with highest sensitivity. It is, however, also important to take assembly length into consideration since the sequence length is critical for target gene-based molecular ecology and diversity analyses. After filtering assemblies with length cutoff of 450 bp, Xander and MegaGTA provided the best sensitivity with all three datasets for *rplB* and *nifH*.

Another aspect of assembly quality is the production of non-target sequences, i.e., false positives. All assemblers produced some, but Xander and MegaGTA by far produced the fewest while GenSeed-HMM, MEGAN, and SAT-Assembler produced the most. Some produced by MEGAN were exceedingly long and matched portions of a genome in the synthetic or mock community with 100% identity. MEGAN assembler works by assembling all reads mapped to a GO (in our case) or KEGG category (Huson et al., 2017). We suspect that the production of non-target contigs has to do with how reads are mapped, and possibly with errors in the mapping file that maps NCBI IDs to functional categories in GO.

In most cases, chimeras are to be expected among close relatives from assembly of shotgun data whether gene-targeted or whole genome. MEGAN is the exception here because, as mentioned above, contigs are usually essentially genes or genome segments of what is in NCBI-nr. Our results are therefore somewhat surprising and encouraging. With the exception of SAT-Assembler, nearly all chimeras detected were between the most closely related sequences suggesting accurate taxonomic classification to the genus level.

"How much sequencing do I need" is often the first question asked when designing a metagenomics project. The answer depends on the target species (usually with specific functions) of interest, since it is difficult to estimate the true diversity (Rodriguez and Konstantinidis, 2014; Rodriguez-R et al., 2018) and also costly to sequence deep enough to cover most species in complex environments (Locey and Lennon, 2016). Therefore, sequencing depth estimates based on a target species or function is critical for experiment planning. With our derived model, the relation between the amount of sequencing data and the probability of assembling a contig with at least "M" bp of the target gene with a size of "S" bp from taxa with a relative abundance of "R" can be determined (**Figure 1**). The relative abundance ("R") can be estimated using common 16s rRNA gene amplicon or qPCR methods. This estimate is a lower bound, since sequencing error, repeats, and micro-heterogeneity among closely related strains could complicate assembly of the target gene.

Because it is difficult to have enough sequencing depth to cover most species in a high diversity sample, follow-up questions are "how many rare members are not assembled" and "how does sequencing depth change the assembled read ratio?" Even though each rare member is only a small percentage of the total community, their sum could be a significant part of the community and thus have a significant role in community function. Missing rare members is an unavoidable problem for all assembly-based methods because there is simply not enough coverage (Guo et al., 2018). There are two cases of rare members: (1) those that are too rare to yield any read coverage and (2) those that have some coverage but not enough to assemble the target gene with minimum length. Here, we focus on the latter. In our soil sample (C1), the number of *rplB* assembled decreased much faster than linear decrease with sequencing depth (**Figure 2**), suggesting that sequencing depth has a strong impact on gene-targeted assembly in diverse communities and thus careful planning on sequencing depth is critical. As an upper bound, the quantity of a targeted gene can be assessed from the number of short reads annotated as the targeted gene without assembly. While this minimizes missing low coverage members, it often includes false positives (low specificity) when there are conserved motifs among protein families. There have been efforts to tackle this problem such as finder function in HMM-GRASPx and ROCKer (Orellana et al., 2017). Also, ROCKer builds gene specific models that set specific sequence similarity score thresholds for different regions of a gene. These kinds of tools can not only improve gene quantification but also could be used as a preprocess step for all above gene-targeted tools, e.g., ROCKer has been shown to improve the accuracy of Xander (Orellana et al., 2017).

All tools reviewed here except MEGAN make use of pHMMs built from reference sequences. The use of pHMMs has clear advantages. It is a faster and more effective way to search gene fragments compared to pairwise alignment as implemented by BLAST or DIAMOND. Additionally, pHMM-based profile search can improve the sensitivity for remotely related protein identification (Eddy, 2009; Zhang et al., 2014; Reyes et al., 2017). The performance of pHMM-based tools, however, is dependent on the quality of the pHMMs used, which in turn is dependent on the appropriateness of the reference sequences used to build them. Ideally, the pHMMs will selectively capture all diversity in the gene family.

The availability of reliable pHMMs may influence the choice of tools used. MEGAN does not require them, and SAT-Assembler is designed to work with pHMMs downloaded from Pfam. Xander (and hence MegaGTA), however, come with a limited set of pHMMs and required reference sequences for finding starting kmers. Instructions are provided for adding capability for additional genes to Xander. The FunGene (Fish et al., 2013) website is provided to help with this task, but knowledge of the gene's diversity is required. Profile HMMs are built to capture conserved regions (domains) of a gene family, and there is usually enough variation to divide the gene family into sub-groups. If the sequences used to build the pHMM do not include all subgroups of the gene, then not all gene diversity will be captured from metagenomic data. In some cases, as was shown for *nosZ* (Sanford et al., 2012), there is too much diversity to be captured by a single pHMM; hence, multiple models are necessary. Based on our experience, if there is large sequence variation in a gene (<50% identity), then it should be split, and subgroups can be defined based their segregation on

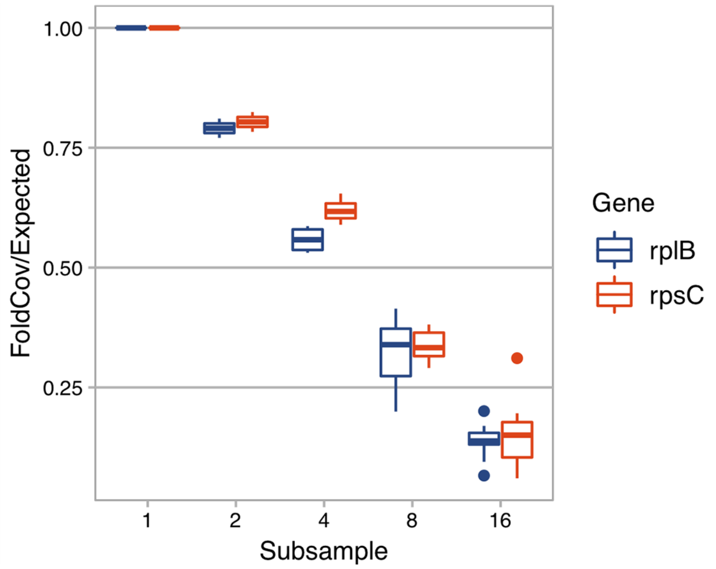


FIGURE 2 | The effect of sequencing depth on the fold coverage of *rplB* or *rpsC* assembled. X axis is the number of subsamples C1 is evenly divided into. Y axis is *rplB* or *rpsC* fold coverage of a subsample divided by expected folded coverage as if it decreases linearly with sequencing depth (the fold coverage of original sample divided by number of even subsamples).

a phylogenetic tree. Thus, results are strongly dependent on the care with which the models are built.

Microbial ecologists are interested in comparing microbiomes among environments or treatments with respect to diversity and function. Metagenomic analyses can answer these questions, but the tools used must accurately assemble and quantify target genes in a manner that allows comparisons among samples. Of the tools reviewed here, only Xander and MegaGTA offer this capability directly (**Table S9**). Their search script includes steps for removing chimeras, clustering reads based on a user-defined distance, providing coverage adjusted counts, and taxonomically matching representative sequences to sequences in a database. An additional script is provided to combine this information from multiple samples to create files that may be imported into phyloseq (McMurdie and Holmes, 2013) as a coverage adjusted OTU table, representative sequences, and, with a function in RDPutils (Quensen, 2018), a corresponding taxonomy table. This gives great flexibility for subsequent analyses. MEGAN can also generate OTU tables and ordinate samples based on taxonomy from all reads, but not in a way that the results are based on a particular set of (pathway related) genes. Additionally, the high proportion of false positives we observed with MEGAN makes using its results for comparative analyses of functional genes questionable. Using SAT-Assembler or GenSeed-HMM results to make like comparisons would require writing additional custom scripts. HMM-GRASPx failed to assemble sequences from complex data, and its authors caution that its results are not quantitative. Most tools except Xander and MegaGTA do not have post-assembly diversity analyses across samples, but they can be improved by applying the post-assembly processing method in Xander. Further improvements can be made on Xander and MegaGTA too. Currently, their post-assembly processing method is designed for assembling each sample individually, but not for pooled assembly, which is common practice applied to increase coverage of rare species. Moreover, they do not directly provide a BIOM table that integrates both OTU table and taxonomy information (McDonald et al., 2012) and can be imported into other commonly used microbial diversity analysis tools such as Mothur (Schloss et al., 2009) and QIIME (Caporaso et al., 2010).

We tested the tools under comparable conditions by using default parameters, which by no means are the optimal parameters; especially kmer or overlap size can strongly impact contig length and number and chimera number. We did not try to find the optimal set of parameters for each tool and only adjusted them when a tool performed significantly more poorly than others, i.e., SAT-Assembler produced too many short and chimeric contigs, and we improved its results by increasing the overlap length.

SUMMARY AND OUTLOOK

Gene-targeted assembly offers advantages for metagenome analysis over whole genome assembly and binning because of (1) higher quality assembly (fewer chimera), (2) more extensive recovery of genes of interest (more sensitivity), and (3) faster and less costly analysis of complex communities which also makes these analyses available to a larger set of researchers. It does, however, give up information on gene context and host taxa that come from genome binning. Long-read sequencing, now available but in its infancy, has the potential to make assembly obsolete, but the present high error rates and low capacity make its reliable and routine use some years away. In the meantime, further improvements of gene-targeted tools, some of which are noted above, will help speed the analysis of the now huge metagenomic data in public databases plus the data from even larger sequencing efforts underway.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR606249 and https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3989263.

AUTHOR CONTRIBUTIONS

JG and JQ performed the analyses under the supervision of JT, JC, YS, QW, and CB. All also helped with the analytical approaches and writing of the manuscript.

FUNDING

Support for this research was provided by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (Awards DE-FC02-07ER64494 and DE-FG02-99ER62848); National Institute of Environmental Health Science, Award N0172U P42ES004911; and by the National Science Foundation, Awards DBI-1356380, DBI-1759892 and DEB 1637653.

ACKNOWLEDGMENTS

We thank Ribosomal Database Project (RDP), Institute for Cyber-Enabled Research (iCER) and High Performance Computing Center (HPCC) at Michigan State University for technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00957/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi: 10.1093/nar/25.17.3389

Alves, J. M. P., de Oliveira, A. L., Sandberg, T. O. M., Moreno-Gallego, J. L., de Toledo, M. A. F., de Moura, E. M. M., et al. (2016). GenSeed-HMM: A tool for progressive assembly using profile HMMs as seeds and its application in Alpavirinae viral discovery from metagenomic data. F*rontiers in Microbiology* 7, 269. doi: 10.3389/fmicb.2016.00269

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G. W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 40 (12), e94. doi: 10.1093/nar/gks251

Awad, S., Irber, L., and Brown, C. T. (2017). Evaluating metagenome assembly on a simple defined community with many strain variants. *bioRxiv* (155358). doi: 10.1101/155358

Brown, C. T., Moritz, D., O'Brien, M., Reidl, F., Reiter, T., and Sullivan, B. (2018). Exploring neighborhoods in large metagenome assembly graphs reveals hidden sequence diversity. *bioRxiv* 462788. doi: 10.1101/462788

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12 (1), 59–60. doi: 10.1038/nmeth.3176

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 (5), 335–336. doi: 10.1038/nmeth.f.303

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome informatics. Intl. Conf. Genome Inf.* 23 (1), 205–211. doi: 10.1142/9781848165632_0019

Edgar, R. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. *bioRxiv* 074252. doi: 10.1101/074252

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 15 (6), 2194–2200. doi: 10.1093/bioinformatics/btr381

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44 (D1), D279–D285. doi: 10.1093/nar/gkv1344

Fish, J. A., Chai, B., Wang, Q., Sun, Y., Brown, C. T., Tiedje, J. M., et al. (2013). FunGene: the functional gene pipeline and repository. *Front. Microbiol.* 4, 291. doi: 10.3389/fmicb.2013.00291

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., and Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl. Environ. Microbiol.* 74 (8), 2461–2470. doi: 10.1128/AEM.02272-07

Gregor, I., Schoenhuth, A., and McHardy, A. C. (2016). Snowball: strain aware gene assembly of metagenomes. *Bioinformatics* 32 (17), 649–657. doi: 10.1093/bioinformatics/btw426

Guo, J. (2018). *rpsC reference database for Xander (Version v1.0) [Dataset]*. doi: 10.5281/zenodo.1410823

Guo, J., Cole, J., Brown, C. T., and Tiedje, J. M. (2018). Comparing faster evolving rplB and rpsC versus SSU rRNA for improved microbial community resolution. *bioRxiv* 435099. doi: 10.1101/435099

Guo, J., Cole, J. R., Zhang, Q., Brown, C. T., and Tiedje, J. M. (2016). Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. *Appl. Environ. Microbiol.* 82 (1), 157–166. doi: 10.1128/AEM.02772-15

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for heuristic determination of minimum cost paths. *Ieee Trans. Syst. Sci. Cybern. SSC* 4 (2), 100–10+. doi: 10.1109/TSSC.1968.300136

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U. S. A.* 111 (13), 4904–4909. doi: 10.1073/pnas.1402564111

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9 (9), 868–877. doi: 10.1101/gr.9.9.868

Huson, D. H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12 (6), e1004957. doi: 10.1371/journal.pcbi.1004957

Huson, D. H., Tappu, R., Bazinet, A. L., Xie, C., Cummings, M. P., Nieselt, K., et al. (2017). Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. *Microbiome* 5, 11. doi: 10.1186/s40168-017-0233-2

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41 (1), e1. doi: 10.1093/nar/gks808

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2 (3), 231–239. doi: 10.1016/0888-7543(88)90007-9

Lawler, E. L. (1972). Procedure for computing K best solutions to discreate optimization problems and its application to shortest path problem. *Manage. Sci. Ser. a-Theory* 18 (7), 401–405. doi: 10.1287/mnsc.18.7.401

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 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, D. H., Huang, Y. K., Leung, C. M., Luo, R. B., Ting, H. F., and Lam, T. W. (2017). MegaGTA: a sensitive and accurate metagenomic gene-targeted assembler using iterative de Bruijn graphs. *BMC Bioinf.* 18 (Suppl 12), 67–78. doi: 10.1186/s12859-017-1825-3

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 (2), 265–272. doi: 10.1101/gr.097261.109

Lippert, R. A., Mobarry, C. M., and Walenz, B. P. (2005). A space-efficient construction of the Burrows-Wheeler transform for genomic data. *J. Comput. Biol.* 12 (7), 943–951. doi: 10.1089/cmb.2005.12.943

Locey, K. J., and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* 113 (21), 5970–5975. doi: 10.1073/pnas.1521291113

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1 (1), 18. doi: 10.1186/2047-217X-1-18

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., et al. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 1, 7. doi: 10.1186/2047-217X-1-7

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *Plos One* 8 (4), e61217. doi: 10.1371/journal.pone.0061217

Myers, E. W. Jr. (2016). A history of DNA sequence assembly. *It-Inf. Technol.* 58 (3), 126–132. doi: 10.1515/itit-2015-0047

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40 (20), e155. doi: 10.1093/nar/gks678

National Research Council (2007). *The new science of metagenomics: revealing the secrets of our microbial planet*. Washington, DC: The National Academies Press.

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15 (6), 387–396. doi: 10.1093/dnares/dsn027

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27 (5), 824–834. doi: 10.1101/gr.213959.116

Orellana, L. H., Rodriguez-R, L. M., and Konstantinidis, K. T. (2017). ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res.* 45 (3), e14. doi: 10.1093/nar/gkw900

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., and Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.* 109 (33), 13272–13277. doi: 10.1073/pnas.1121464109

Peltola, H., Soderlund, H., and Ukkonen, E. (1984). SEQAID—a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Res.* 12 (1), 307–321. doi: 10.1093/nar/12.1Part1.307

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L., (2010). "IDBA—a practical iterative de Bruijn graph de novo assembler," in *Research in Computational Molecular Biology, Proceedings*. Ed. Berger, B. (Berlin: Springer-Verlag), 426–440. doi: 10.1007/978-3-642-12683-3_28

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11), 1420-1428. doi: 10.1093/bioinformatics/bts174

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 (7285), 59–U70. doi: 10.1038/nature08821

Quensen, J. (2018). "RDPutils: R Utilities for processing RDPTool output," in *R package version 1.4.1 ed*. https://github.com/jfq3/RDPutils.

Reyes, A., Alves, J., Durham, A., and Gruber, A. (2017). Use of profile hidden Markov models in viral discovery: current insights. *Adv. Genomics Genet.* 7, 29–45. doi: 10.2147/AGG.S136574

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38 (20), e191. doi: 10.1093/nar/gkq747

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018). Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *Msystems* 3 (3), e00039-18. doi: 10.1128/mSystems.00039-18

Rodriguez, R. L., and Konstantinidis, K. T. (2014). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30 (5), 629–635. doi: 10.1093/bioinformatics/btt584

Sanford, R. A., Wagner, D. D., Wu, Q., Chee-Sanford, J. C., Thomas, S. H., Cruz-Garcia, C., et al. (2012). Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proc. Natl. Acad. Sci. U. S. A.* 109 (48), 19709–19714. doi: 10.1073/pnas.1211238109

Schaeffer, R. D., Liao, Y., Cheng, H., and Grishin, N. V. (2017). ECOD: new developments in the evolutionary classification of domains. *Nucleic Acids Res.* 45 (D1), D296–D302. doi: 10.1093/nar/gkw1137

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 (23), 7537–7541. doi: 10.1128/AEM.01541-09

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods* 14 (11), 1063–106+. doi: 10.1038/nmeth.4458

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* 15 (6), 1882–1899. doi: 10.1111/1462-2920.12086

Simpson, J. T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 22 (3), 549–556. doi: 10.1101/gr.126953.111

Simpson, J. T., and Pop, M. (2015). The theory and practice of genome sequence assembly. *Annu. Rev. Genomics Hum. Genet*. 16 (1), 153–172. doi: 10.1146/annurev-genom-090314-050032

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19 (6), 1117–1123. doi: 10.1101/gr.089532.108

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348 (6237), 1261359. doi: 10.1126/science.1261359

Wang, Q., Fish, J. A., Gilman, M., Sun, Y., Brown, C. T., Tiedje, J. M., et al. (2015). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 3 (1), 32. doi: 10.1186/s40168-015-0093-6

Yang, Y., Zhong, C., and Yooseph, S. (2015). SFA-SPA: a suffix array based short peptide assembler for metagenomic data. *Bioinformatics* 31 (11), 1833–1835. doi: 10.1093/bioinformatics/btv052

Yen, J. Y. (1971). Finding th K shortest loopless paths in a network. *Manage. Sci. Ser. a-Theory* 17 (11), 712–716. doi: 10.1287/mnsc.17.11.712

Zhang, Y., Sun, Y., and Cole, J. R. (2014). A scalable and accurate targeted gene assembly tool (SAT-Assembler) for next-generation sequencing data. *Plos Comput. Biol.* 10 (8), e1003737. doi: 10.1371/journal.pcbi.1003737

Zhong, C., Yang, Y., and Yooseph, S. (2016). GRASPx: efficient homolog-search of short peptide metagenome database through simultaneous alignment and assembly. *BMC Bioinf.* 17 (Suppl 8): 283. doi: 10.1186/s12859-016-1119-1

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

published: 17 October 2019
doi: 10.3389/fgene.2019.01022

Embracing Ambiguity in the Taxonomic Classification of Microbiome Sequencing Data

Nidhi Shah[1,2,3], *Jacquelyn S. Meisel*[1,2,3,4] *and Mihai Pop*[1,2,3,4]*

[1] Department of Computer Science, University of Maryland, College Park, College Park, MD, United States, [2] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, College Park, MD, United States, [3] University of Maryland Institute for Advanced Computer Studies, College Park, MD, United States, [4] Center for Health-related Informatics and Bioimaging, University of Maryland, College Park, College Park, MD, United States

OPEN ACCESS

Edited by:
Joao Carlos Setubal,
University of São Paulo, Brazil

Reviewed by:
Graziano Pesole,
University of Bari Aldo Moro,
Italy
Erik Scott Wright,
University of Pittsburgh,
United States

***Correspondence:**
Mihai Pop
mpop@umiacs.umd.edu

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

Received: *23 April 2019*
Accepted: *24 September 2019*
Published: *17 October 2019*

Citation:
Shah N, Meisel JS and Pop M (2019) Embracing Ambiguity in the Taxonomic Classification of Microbiome Sequencing Data. Front. Genet. 10:1022. doi: 10.3389/fgene.2019.01022

The advent of high throughput sequencing has enabled in-depth characterization of human and environmental microbiomes. Determining the taxonomic origin of microbial sequences is one of the first, and frequently only, analysis performed on microbiome samples. Substantial research has focused on the development of methods for taxonomic annotation, often making trade-offs in computational efficiency and classification accuracy. A side-effect of these efforts has been a reexamination of the bacterial taxonomy itself. Taxonomies developed prior to the genomic revolution captured complex relationships between organisms that went beyond uniform taxonomic levels such as species, genus, and family. Driven in part by the need to simplify computational workflows, the bacterial taxonomies used most commonly today have been regularized to fit within a standard seven taxonomic levels. Consequently, modern analyses of microbial communities are relatively coarse-grained. Few methods make classifications below the genus level, impacting our ability to capture biologically relevant signals. Here, we present ATLAS, a novel strategy for taxonomic annotation that uses significant outliers within database search results to group sequences in the database into partitions. These partitions capture the extent of taxonomic ambiguity within the classification of a sample. The ATLAS pipeline can be found on GitHub [https://github.com/shahnidhi/outlier_in_BLAST_hits]. We demonstrate that ATLAS provides similar annotations to phylogenetic placement methods, but with higher computational efficiency. When applied to human microbiome data, ATLAS is able to identify previously characterized taxonomic groupings, such as those in the class *Clostridia* and the genus *Bacillus*. Furthermore, the majority of partitions identified by ATLAS are at the subgenus level, replacing higher-level annotations with specific groups of species. These more precise partitions improve our detection power in determining differential abundance in microbiome association studies.

Keywords: microbiome, taxonomy, classification, *16S rRNA* marker gene, high-throughput sequencing

INTRODUCTION

The microbiome plays an important role in human and ecological health. One of the first steps in microbial characterization is taxonomic classification. Modern taxonomy was founded in the 1750s by Swedish botanist Carl Linnaeus, who worked to establish a hierarchical classification of organisms based on shared characteristics that were consistent and universally accepted. While the initial taxonomy was able to capture the complex relationships between organisms, maintaining and expanding this taxonomy remain a challenge (Godfray, 2002). In particular, the microbial taxonomy has significantly evolved since the time of Linnaeus, most notably with the advent of next-generation sequencing technologies that enable us to examine microbiota with greater resolution.

Many microbiome studies involve extracting DNA from a microbial community and amplifying and sequencing the *16S rRNA* gene, a gene encoding part of the ribosomal complex. This gene is highly conserved across prokaryotes and can be amplified even from previously unknown organisms. Originally, phylogenetic approaches (Yang and Rannala, 2012) were used to build trees to relate organisms based on how they evolved from each other. These trees were independent of taxonomic annotation and were instead generated directly from sequencing data *via* neighbor-joining (Zhang and Sun, 2008), maximum parsimony (Fitch, 1971; Tamura et al., 2011), maximum likelihood (Stamatakis, 2006), or other methods. Because building a phylogenetic tree is computationally expensive, we often perform taxonomic annotation by searching against a reference database of "known" sequences instead.

There are several limitations to nonphylogenetic approaches. First, it is often impossible to obtain confident species- or even genus-level classifications within samples due to the lack of discriminative power of the sequenced marker gene (Barb et al., 2016). The *16S rRNA* gene contains nine taxonomically discriminating hypervariable regions, however, there is no single hypervariable region of the gene that can distinguish between all species. Additionally, reference databases are not always representative of a sample and are dominated by a small subset of easy to isolate organisms found at higher abundances (Walker et al., 2014). Sequencing data in reference databases is largely biased toward pathogenic microbes and organisms commonly found in developed countries. The organisms found in many studies (e.g., in environmental communities or in developing countries) have no near neighbors in reference databases, making it difficult to assign to them accurate taxonomic labels.

Another problem with modern analysis of microbial communities is the relatively coarse-grained resolution obtained, which limits our ability to capture biologically relevant signals. This stems from the need to simplify computational workflows. Most classification algorithms utilize just seven taxonomic levels and often ignore intermediate taxonomic ranks. This problem is further compounded by errors and missing information in databases, as well as inherent ambiguities in the taxonomic assignment of some sequences. Some taxonomic ambiguity may also arise by taxonomic mislabeling of some entries in the database. Current software tools frequently rely on "most recent common ancestor" (MRCA) strategies to provide an annotation at the most general taxonomic level that encompasses all of the possible annotations of a sequence. As a result, few methods ever make classifications below the genus level, and, frequently, sequences are only classified at the family, class, or even phylum level.

As the number and size of sequencing datasets continues to grow, taxonomic classification methods often make trade-offs between speed and accuracy. Different tools have been developed for taxonomic annotation, using either composition-based, sequence-similarity, or phylogenetic-placement methods (Altschul et al., 1990; Liu et al., 2011; Nguyen et al., 2014; Wood and Salzberg, 2014; Ounit et al., 2015). Composition based and sequence-similarity based approaches are fast and require less computational power, but only work well when the microorganisms in the sample have near neighbors in the database. On the other hand, phylogenetic-placement based methods statistically model the evolutionary processes that generate the query sequences and are computationally expensive, but allow classification even if only distant neighbors are found in databases.

Here, we propose a novel strategy for taxonomic annotation that adequately captures and represents the complexity of the bacterial world, providing more specific and more interpretable characterizations of the composition of microbial communities while also capturing the inherent ambiguity in the classification of sequences. Our strategy is sequence-similarity based and builds upon our recent work on detecting significant "outliers" within database search results (Shah et al., 2018), allowing us to characterize, in a sample-specific manner, the extent of taxonomic ambiguity within the classification. In this work, detecting "outliers" refers to separating the phylogenetically most closely related BLAST matches from matches to sequences from more distantly related organisms. This approach allows us to make assignments at the species level, and even when such assignment is not possible, we may be able to identify the few species within a genus that are the most likely origin of the fragment being analyzed. Such information is particularly relevant in clinical applications, allowing us to distinguish between the pathogenic and nonpathogenic members of the same genus even if the specific species cannot be uniquely identified. It is also important to stress that, by design, our method is conservative - it only provides a classification, even at an intermediate taxonomic level, only when it has high confidence that such a classification is supported by the data. In some cases, particularly for genes such as the*16S rRNA*, which have poor discriminatory power within certain taxonomic group, this will result in sequences being left unclassified, or only classified at high taxonomic levels.

Our method, called "ATLAS-**A**mbiguous **T**axonomy e**L**ucidation by **A**pportionment of **S**equences," is implemented in Python and released under the open-source MIT license on GitHub [https://github.com/shahnidhi/outlier_in_BLAST_hits]. ATLAS supplements sequence-similarity based approaches with a graph-based approach to identify and group sequences with ambiguous database assignments. We demonstrate that ATLAS yields similar results to phylogenetic methods, but with reduced computational requirements. We use ATLAS to reexamine over 2000 samples from the Human Microbiome Project (HMP) (The Human Microbiome Project Consortium, 2012) and interrogate almost one-thousand stool samples from the Global Enteric Multicenter Study (GEMS) of young children in low-income countries with moderate-to-severe diarrhea (Pop et al., 2014). The HMP dataset provides a large sample size of short-read sequencing data, and the GEMS data is from a population that is underrepresented in our current genomic databases and contains a large proportion of uncharacterized organisms. In these datasets, we identify partitions matching previously defined groupings of organisms within the *Bacillus* genus and the *Clostridia* class. We also demonstrate that the partitions identified by ATLAS increase the power of differential abundance analyses. Although our results specifically focus on data from *16S rRNA* gene surveys, ATLAS can be used with any marker gene sequencing data to characterize the taxonomic composition of a microbial community and to determine microbiome associations with human and ecological health.

MATERIALS AND METHODS

ATLAS Algorithm Overview

ATLAS groups sequences into biologically meaningful taxonomic partitions by querying them against a reference database and identifying and clustering significant database hits. ATLAS has two phases (see **Figure 1**): (i) identifying significant database hits for query sequences and (ii) generating database partitions (clusters) that capture the ambiguity in the assignment process.

Aligning Query Sequences and Identifying Significant Database Hits

ATLAS uses BLAST (Altschul et al., 1990) to align each sequence in an input set of uncharacterized query sequences to sequences in a reference set (using parameters *-outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq"*). The previously published "BLAST outlier detection" algorithm is used to identify significant top BLAST hits for each query sequence (Shah et al., 2018). We refer to these BLAST hits as outliers. In brief, the "BLAST outlier detection" algorithm constructs a multiple sequence alignment of the query sequence and the top BLAST hits from the BLAST-generated pairwise alignments. It then uses the Bayesian integral log odds (BILD) score (Brown et al., 1993; Altschul et al., 2010) to determine whether the multiple alignment can be split into two groups

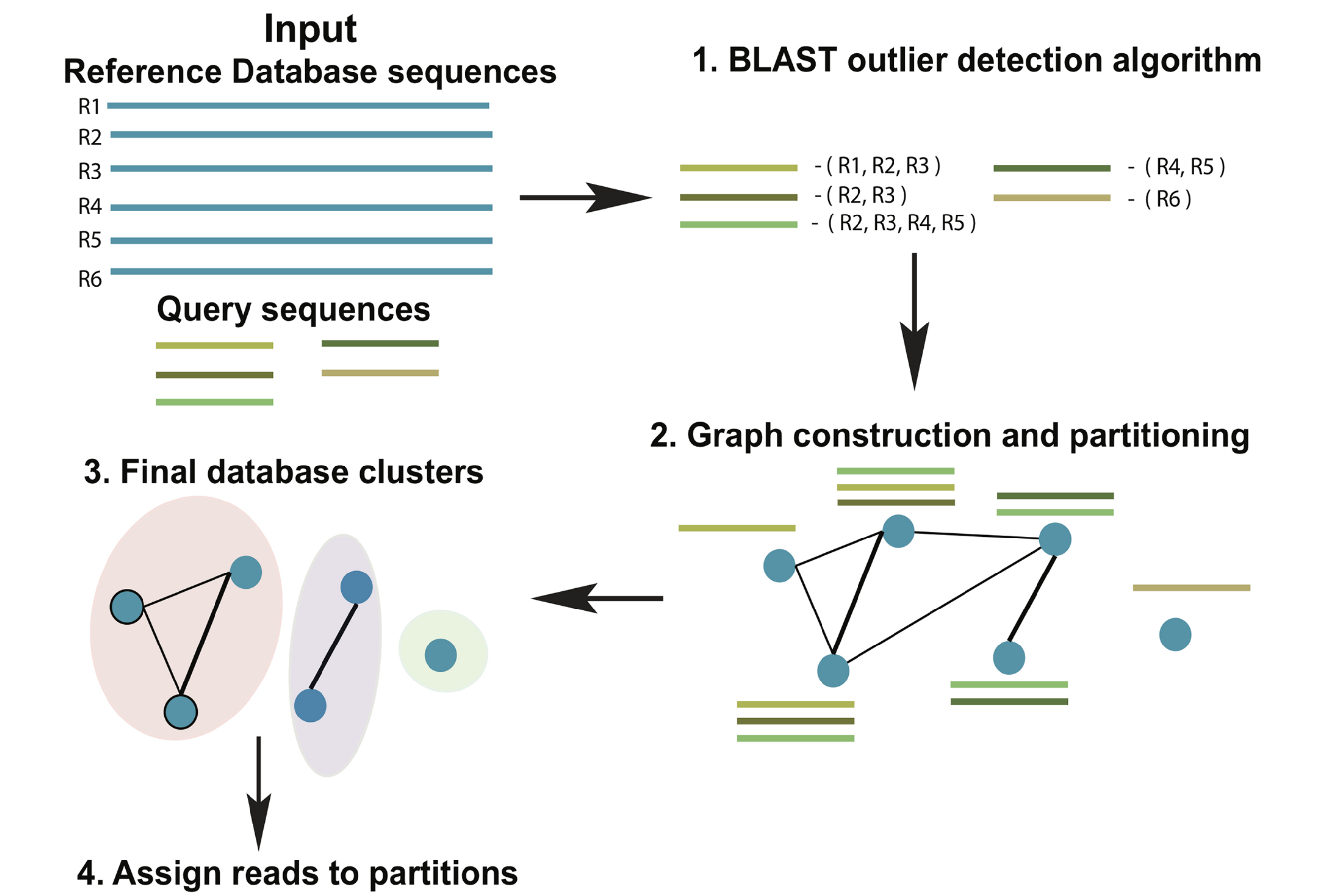

Input
Reference Database sequences
R1
R2
R3
R4
R5
R6
Query sequences
1. BLAST outlier detection algorithm
- (R1, R2, R3)
- (R2, R3)
- (R2, R3, R4, R5)
- (R4, R5)
- (R6)
2. Graph construction and partitioning
3. Final database clusters
4. Assign reads to partitions

FIGURE 1 | Schematic diagram of the ATLAS pipeline. ATLAS takes in query sequences from a marker gene and searches them against a reference database to identify outlier sequences. It then constructs a graph of database sequences and clusters those that are commonly identified together into partitions.

that model the data better than a single group. This process identifies which BLAST hits are significantly associated with the query sequence, without resorting to *ad hoc* cut-offs on percent identity, bit-score, and/or E-value.

Generating Database Partitions That Capture the Ambiguity in the Assignment Process

Ambiguity in the taxonomic assignment process occurs for two main reasons. First, the query sequence may not have any near-neighbors in the database, resulting in multiple equally-good hits (neighbors) (**Figure 2**). Second, the query sequence may align to a genomic region that is conserved across distantly related organisms. Our method characterizes this ambiguity in a sample-specific manner, identifying database sequences that are equivalent with respect to their similarity to the set of query sequences.

From all query sequences and their set of related database sequences (outlier set), we construct a confusion graph. The nodes in the graph represent sequences in the database, whereas the edges link nodes that are present together in the outlier set for at least one query sequence. The edges are weighted by the number of query sequences that shares the same nodes (reference database sequences) within the outlier set. Tightly-knit subcommunities in the confusion graph indicate database sequences that are equivalent based on similarity to the set of query sequences, and hence, should be clustered together. To identify these subcommunities, we remove all the low-weight edges (below mean – 2 * std.dev of all edge weights) and identify strong communities in the network using the Louvain community detection algorithm, which optimizes the modularity of the network (Blondel et al., 2008). These subcommunities become the final database partitions (clusters). ATLAS partitions can be singletons (consist of one reference database sequence).

Assigning Query Sequences to the Partitions

A query sequence is assigned to a database partition if a certain percentage (user-defined, default 50%) of the database sequences in the outlier set belong to the partition. ATLAS does not classify the query sequence if no BLAST outliers can be detected, or the query sequence does not meet these thresholds.

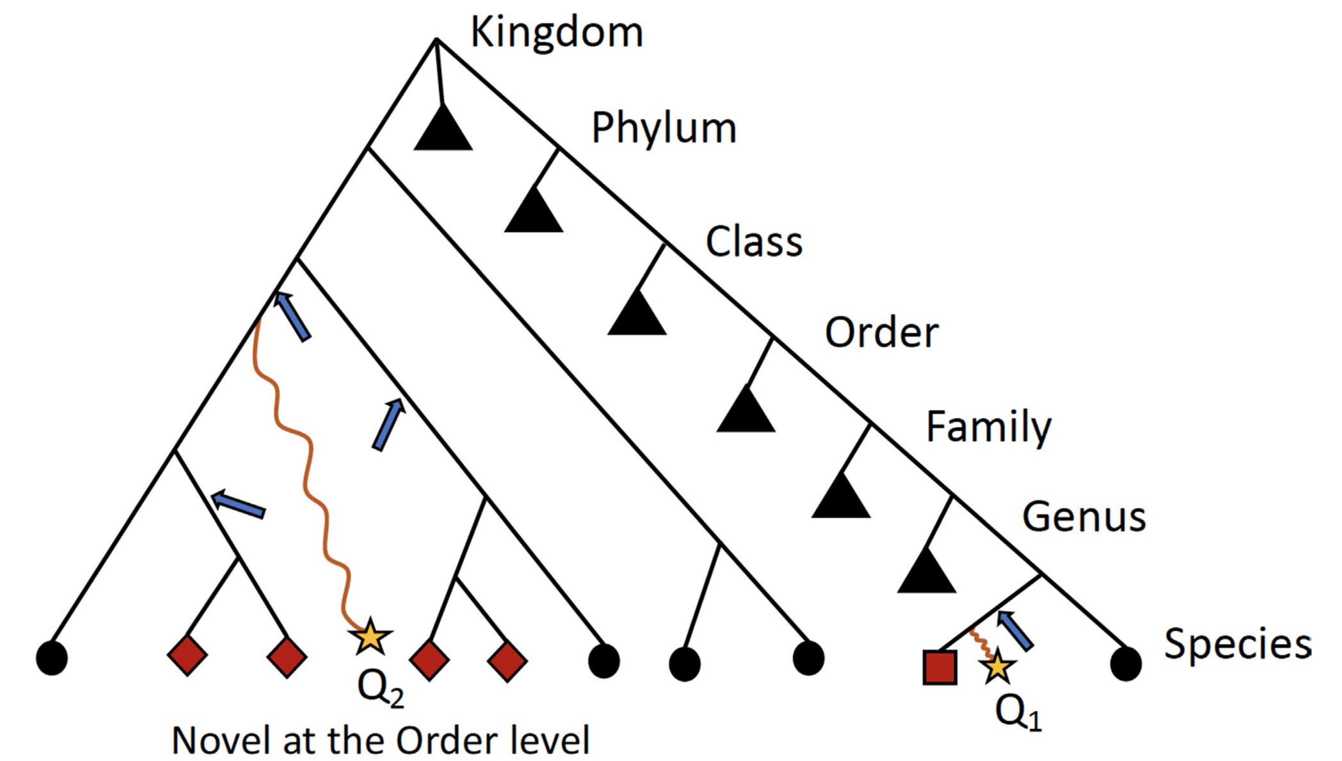


FIGURE 2 | Schematic detailing when ATLAS will provide the greatest improvement to taxonomic annotation. Shown is a simple example of a phylogenetic tree with taxonomic information of reference sequences, where the leaves are actual sequences in the database. When a query sequence (yellow stars) has near neighbors in the reference, such as Q_1, most algorithms will be able to correctly classify the sequence. However, if a sequence, such as Q_2, does not have many near neighbors in the database, computationally expensive phylogenetic methods are required for accurate placement (blue arrows) and annotation. ATLAS captures groups (or partitions) of database sequences (red nodes) that are commonly confused during the annotation process and assigns them to the query sequence (square node for Q_1 and diamond nodes for Q_2). Black triangles show collapsed portion of the tree. While this schematic is overly simplified and real phylogenies are much more complex, this is illustrating that ATLAS will provide additional information when query sequences do not have near neighbors in the database. This represents ideal cases, where 16S rRNA phylogeny and taxonomic annotations are congruent.

The goal of ATLAS is only to classify sequences when it has enough confidence in the taxonomic assignment. Sequences that remain unclassified by ATLAS should be further examined with more sophisticated approaches, such as phylogenetic placement methods. For each query sequence, ATLAS provides a species list based on the reference database sequences included within the assigned partition. To provide a high-level summary of the data and simplify the comparison to other annotation methods, ATLAS also assigns to query sequences the MRCA of all sequences belonging to a partition. These partitions of database sequences attempt to capture the most accurate granularity of taxonomic assignment without relying solely on the main taxonomic levels.

Comparison to Other Taxonomic Assignment Methods

To benchmark ATLAS with other widely used taxonomic annotation methods, we downloaded TAXXI test and train datasets (sp_ten_16s_v35) from a recent study that benchmarked taxonomic methods for microbiome studies (Edgar, 2018). We compared ATLAS with RDP classifier (Wang et al., 2007), mothur (Schloss et al., 2009), UCLUST (Edgar, 2010), SortMeRNA (Kopylova et al., 2012), and the top BLAST hit. RDP classifier, mothur, and UCLUST were run with 80% confidence threshold. All methods except ATLAS were run *via* QIIME v. 1.9.1 (Caporaso et al., 2010), using the script assign_taxonomy.py. Metrics for method comparison were calculated as previously published (Edgar, 2018).

We also compared ATLAS to the phylogenetic placement method, TIPP. We ran TIPP with the 16S rRNA reference package (rdp_bacteria.refpkg) provided by the authors (https://github.com/tandyw/tipp-reference/releases/download/v2.0.0/tipp.zip). We used the alignment subset size of 100 and the placement subset size of 1,000, and the default values for alignment and placement thresholds.

Analysis of Samples From the Human Microbiome Project (HMP)

The OTU table and representative sequence FASTA files for the V1-V3 hypervariable region of the *16S rRNA* gene sequenced as part of the Human Microbiome Project (The Human Microbiome Project Consortium, 2012) were downloaded from https://www.hmpdacc.org/HMQCP/. We used the 16S rRNA reference package from TIPP for ATLAS and ran it with default settings. The OTU table was filtered to retain OTUs with at least 20 reads and samples containing at least 1,000 reads.

Analysis of Samples From the GEMS Study of Diarrheal Disease

A total of 992 samples were analyzed from a previously published study of diarrheal disease in children in low-income countries that sequenced the V1-V2 region of the *16S rRNA* gene (Pop et al., 2014). In this study, moderate-to-severe diarrhea cases were compared to age- and gender-matched healthy controls. Data was downloaded *via* Bioconductor, using the msd16s package. We used the 16S rRNA reference package from TIPP for ATLAS and ran it with default settings. The dataset was filtered to retain only OTUs with at least 20 reads total and found in at least 10% of case or 10% of control samples.

Significantly differentially abundant OTUs were identified between cases and controls using the R package metagenomeSeq (Paulson et al., 2013), accounting for age in months, country, and sample read counts as potential confounding factors. OTUs were also aggregated separately by genus and by partition. Significant findings were reported for features that had fold change or odds ratio exceeding 2 in either cases or controls and a significant statistical association ($P < 0.05$) after Benjamini-Hochberg correction for multiple testing.

Analysis of Samples From Bangladeshi Children With Acute Diarrhea

A total of 142 samples were analyzed from a previously published study of acute diarrhea in Bangladeshi children that sequenced the V3-V4 region of the *16S rRNA* gene (Kieser et al., 2018). Fastq files were downloaded from BioProject SRP119744, using the SRA toolkit v. 2.8.2 and processed in QIIME v. 1.9.1. We used the 16S rRNA reference package from TIPP for ATLAS and ran it with default settings, identifying 77 partitions.

RESULTS

ATLAS Captures Similar Information as Phylogenetic Placement Algorithms

We compared the taxonomic assignments generated by ATLAS for the HMP and GEMS datasets to the labels generated by TIPP (Nguyen et al., 2014). Because TIPP relies on a phylogenetic approach for taxonomic annotation, it accounts for evolutionary divergence and, therefore, can more effectively analyze sequences without near neighbors in the database than non-phylogenetic methods. We assume here that the classifications provided by TIPP are most accurate because the ground-truth is not available for real datasets. The taxonomic assignments made by ATLAS and TIPP showed 97% and 98% agreement with TIPP assignments at the genus level for GEMS and HMP datasets, respectively (**Figures 3A**, **B**). Importantly, when TIPP could confidently assign a species level classification label to a query sequence, but ATLAS could not, the partition assigned by ATLAS for the majority of query sequences contained the species assigned by TIPP (**Table 1**). The algorithm used by TIPP identifies multiple putative placements of a sequence within the backbone tree representing the reference database. In the vast majority of cases, the partitions identified by ATLAS contained the database sequences selected by TIPP (**Supplemental Figure 1**). Compared to TIPP, ATLAS had a lower run time and only added a small overhead to the run time of BLAST (**Figure 3C**).

We also compared ATLAS to nonphylogenetic approaches (**Supplemental Figure 2**) on the sp_ten_16s_v35 TAXXI benchmarking dataset where the ground truth is known (Edgar, 2018). Compared to other methods, ATLAS has similar or better overclassification and misclassification rates at all taxonomic levels. However, ATLAS often has a higher underclassification

TABLE 1 | Comparison between our approach (ATLAS) and a phylogenetic method (TIPP) examining species level assignments. For most query sequences ATLAS assigned partition contains group of species, as it is often impossible to get species-level resolution. Here, we compare how ATLAS performs when TIPP provides species-level classification.

		GEMS	HMP
A.	Number of query sequences classified by TIPP at the species level	13,050	10,086
	Number of query sequences assigned to a partition that contained TIPP's species	12,847	8,999
B.	Number of query sequences classified at species level by ATLAS that match TIPP's labeling	29	128
	Number of query sequences classified at species level by ATLAS that did not match TIPP's labeling	0	85
	Number of query sequences classified at species level by ATLAS but not by TIPP	18	36

(A) For query sequences where ATLAS partitions do not have a species-level MRCA, the assigned partition contains reference sequences that match TIPP's assigned species. (B) For query sequences where ATLAS partitions do have a species-level MRCA, many of the assigned partitions match TIPP's classification.

rate, particularly at lower taxonomic ranks. This behavior is intentional as ATLAS is meant to serve as a first-level analysis, followed by more sophisticated approaches (such as phylogenetic placement) for the sequences that cannot be confidently classified through sequence similarity searches.

Relationship Between ATLAS Partitions and Standard Taxonomic Levels

ATLAS grouped OTU representative sequences into 185 and 109 non-singleton partitions in the HMP and GEMS datasets, respectively (**Table 2**). A large number of these partitions each have an MRCA at the genus level, suggesting that they are capturing sub-genus information (**Figure 4**). Often, there is not enough information encoded in the short *16S rRNA* gene sequence to offer species-level resolution. However, ATLAS is able to group similar species within a genus, providing resolution that is more specific than the genus level. For instance, in the HMP data, ATLAS identified seven partitions belonging to the genus *Bacillus*

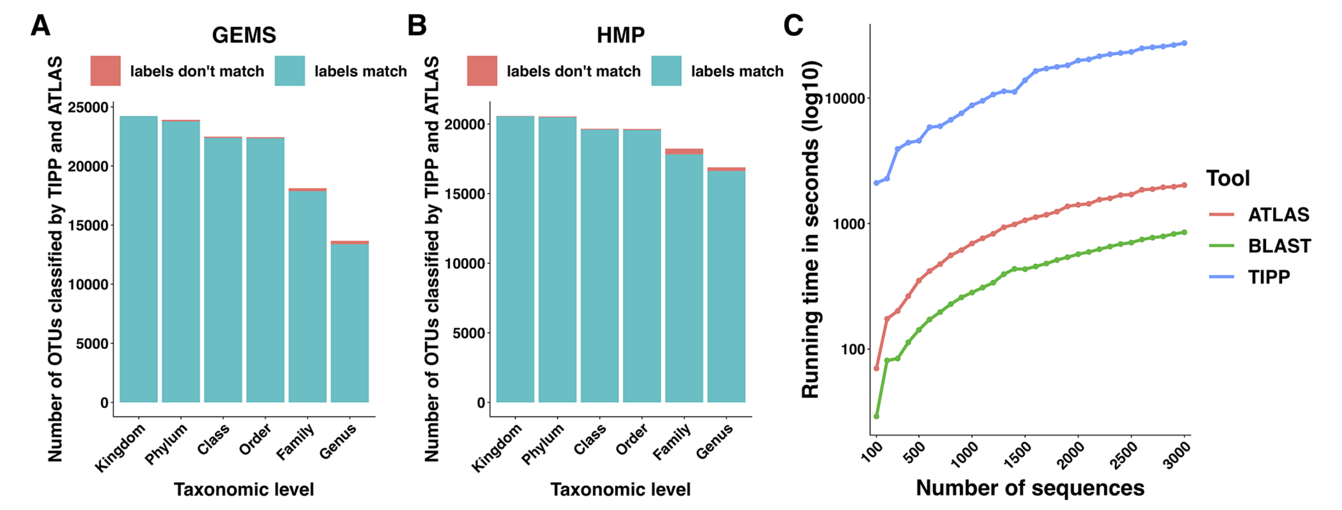


FIGURE 3 | ATLAS generates classifications similar to phylogenetic placement methods at an improved speed. Taxonomic labels assigned by TIPP and ATLAS agree at all taxonomic levels for both **(A)** GEMS and **(B)** HMP datasets. **(C)** The ATLAS pipeline adds minimal post-processing time (in seconds) to standard BLAST analyses, but significantly outperforms TIPP.

TABLE 2 | Number of OTUs and partitions in the HMP and GEMS datasets pre and postfiltering.

	HMP		GEMS		
	OTU	**Partition**	**OTU**	**Genus**	**Partition**
Sequencing Technology	Illumina V1-V3		454 V1-V2		
Number of Samples Post Filtering	2,711 180 gut, 1,553 oral, 719 skin, 259 vagina		992 508 Cases, 484 Controls		
Number of Features Pre-Filtering	43,140 OTUs	307 partitions and 22,578 non-partitioned OTUs	26,044 OTUs	172 genera	122 partitions and 1,819 non-partitioned OTUs
Number of Features Post-Filtering	36,560 OTUs	257 partitions and 17,819 non-partitioned OTUs	10,774 OTUs	149 genera	112 partitions and 924 non-partitioned OTUs

Samples with >1,000 reads were retained for analysis. In the HMP data, features were retained if they had at least 20 total reads or were found in at least 5 samples. In the GEMS data, features were retained if they had at least 20 total reads or were found in at least 10% of case or control samples.

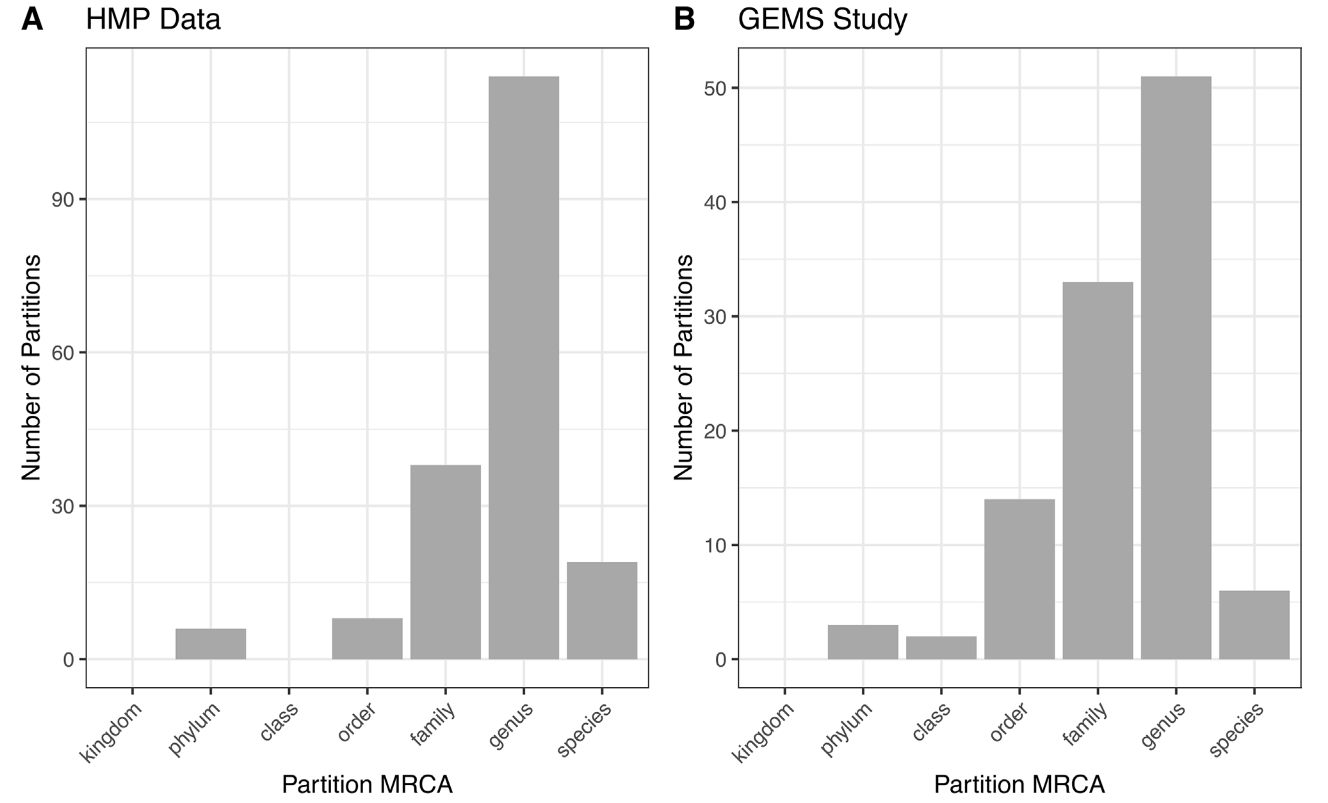


FIGURE 4 | ATLAS partitions for HMP and GEMS data typically capture subgenera information. Most partitions have the most recent common ancestor at the genus level for both **(A)** HMP and **(B)** GEMS datasets.

(**Supplemental Figure 3**). Importantly, reference sequences in partition 156 capture members of the *Bacillus cereus* species group, including *B. cereus*, *B. thuringiensis*, *B. mycoides*, and *B. weihenstephanensis* (Liu et al., 2015). These species have very high sequence similarity and have been shown to play significant roles in human and environmental health (Rasko et al., 2005). ATLAS partition 121 corresponds to the *Bacillus subtilis* group, including species such as *B. subtilis*, *B. licheniformis*, and *B. amyloliquefaciens* (Bhandari et al., 2013). Given the diverse function and pathogenic potential of species within this genus, the distinction of these two groups provides additional benefit to microbiome analyses.

It is important to note that ATLAS partitions are derived purely from sequence similarity; they do not take into consideration any taxonomic or phylogenetic information. Given our incomplete knowledge of microbial diversity and the inherent limitations of 16S rRNA sequences for taxonomic classification, these sub-genus partitions should be further examined and validated.

The percentage of query sequences assigned to partitions spanning multiple genera was 8% for the HMP data and 39% for the GEMS data. Some of these higher-level partition groupings reflect limitations in the hypervariable region of the *16S rRNA* gene sequenced. For instance, in both the HMP and GEMS data, ATLAS identified a single partition spanning the *Enterobacteriaceae* family. While it would be beneficial to distinguish between *Escherichia* and *Shigella* species in the GEMS dataset, the V1-V2 and V1-V3 hypervariable regions of the *16S rRNA* marker gene are insufficient for discrimination (Chakravorty et al., 2007).

Other partitions with higher-level MRCA capture established phylogenetic groupings that span multiple genera. ATLAS was able to capture well-known phylogenetic groupings in the class *Clostridia* (Collins et al., 1994; Johnson and Francis, 1975). In the GEMS data, ATLAS identified 15 partitions comprising sequences from the *Clostridia* class. Of particular note, partition 84 contains *Acetobacterium* species in Clostridial group XV, partition 81 contains members of Clostridial group XI, and Clostridial group I is represented in partitions 5 and 6 (**Supplemental Figure 4**). Clostridial groups encompassed by partitions 0, 81, and 84 contained multiple genera, highlighting the utility of using partitions based on information from the sequences themselves rather than solely relying on modern taxonomic groupings. Interestingly, eight of these partitions were significantly differentially enriched in healthy control samples, supporting the role of *Clostridia* in the maintenance of gut homeostasis (Lopetuso et al., 2013).

ATLAS Partitions Improve the Power of Microbiome-Disease Association Studies

We explored whether ATLAS partitions could provide improved resolution over OTUs in differential abundance analyses. The original GEMS dataset contains 26,044 OTUs, many of which are not prevalent or abundant enough to provide statistical power for identifying associations between health and disease. Filtering OTUs and partitions according to their abundance and prevalence, we retained just those that contained at least 20 sequences and were found in at least 10% of the samples. Only 10,774 OTUs, comprising just 41% of the sequences in the dataset, were retained, whereas ATLAS partitions retained after filtering contained 25,135 total OTUs, comprising 97% of the sequences in the dataset (**Table 2**).

We identified statistically significantly different features between cases with diarrheal disease and healthy controls (**Table 3**). We performed this analysis separately on (i) OTUs, (ii) OTUs aggregated by genus-level assignments, and (iii) OTUs aggregated by ATLAS partitions. Compared to the OTU analysis, OTUs aggregated at the genus-level generally identified more significant OTUs, but fewer overall significant dataset sequences. This is potentially impacted by the fact that 2,411 OTUs and 899,322 sequences had no assignment at the genus level. OTUs aggregated by ATLAS partitions identified a greater number of significant OTUs and sequences enriched in the control samples. When looking at the 10,774 OTUs included in both the OTU-level and partition-based analyses, the majority agreed on differential abundance results (i.e., they were significant or not significant in both analyses) (**Table 4**). Forty-one percent were significant by the partition analysis, but not by OTU based methods. These OTUs were most likely lower abundant community members that became significant as they were aggregated with similar, more abundant microbiota. The few remaining OTUs were significant at the OTU level but not in our partition-based analyses and generally belonged to low abundance genera (**Supplemental Figure 5**).

TABLE 3 | Number of OTUs, genera, and ATLAS partitions that are statistically significantly different between moderate-to-severe diarrheal cases and healthy controls.

	OTU	Genus	Partition
Significant Features with increased expression in case samples	679 OTUs (415,257 sequences)	16 genera (892 OTUs, 342,960 sequences)	13 partitions and 71 non-partitioned OTUs (692 OTUs, 189,005 sequences)
Significant Features with increased expression in control samples	1,112 OTUs (637,591 sequences)	22 genera (1,626 OTUs, 447,680 sequences)	17 partitions and 108 non-partitioned OTUs (4,917 OTUs, 1,300,544 sequences)
Non-significant Features	8,983 OTUs (2,448,992 sequences)	105 genera (5,845 OTUs, 1,811,878 sequences)	77 partitions and 745 non-partitioned OTUs (5,165 OTUs, 2,012,291 sequences)

Features generated from 3,501,840 GEMS dataset sequences were considered differentially abundant if they had a fold change or odds ratio exceeding 2 in either cases or controls and the statistical association was significant (P < 0.05) after Benjamini-Hochberg correction for multiple testing. Singleton partitions have a single OTU mapped to them. Note that when aggregating at the genus level, 2,411 OTUs and 899,322 sequences had no assignment.

TABLE 4 | Confusion matrix highlighting the number of shared/unshared statistically significant OTUs and ATLAS partitions.

		OTUs	
		Not Significant	**Significant**
Partitions	**Not Significant**	4,557	608
	Significant	4,426	1,183

Features were considered differentially abundant between healthy controls and diarrheal cases if they had a fold change or odds ratio exceeding 2 in either cases or controls and the statistical association was significant (P < 0.05) after Benjamini-Hochberg correction for multiple testing.

We also applied ATLAS to a separate acute diarrhea dataset from children in Bangladesh (Kieser et al., 2018), which used a different hypervariable region of the *16S rRNA* gene, a different sequencing platform, and different downstream analyses. Within this dataset, we also identified sub-genus level partitions **(Supplemental Figure 6A).** Many of the sub-genus level partitions in the Bangladesh dataset were in *Lactobacillus*, *Streptococcus, Helicobacter,* and *Campylobacter*, genera which are commonly associated with diarrheal disease **(Supplemental Figure 6B).**

DISCUSSION

As DNA sequencing technologies become faster and cheaper, the number of microbiome studies are rapidly increasing. These studies are aimed at both developing a better understanding of the microbial communities inhabiting the world and at characterizing the association between microbiota and health. Accurate taxonomic assignment is a critical requirement for the interpretation of the data generated in such studies. Current approaches for taxonomic annotation fall at two extremes – computationally intensive phylogenetic inference methods that can accurately classify even sequences that are only distantly related to the reference database and fast approaches based on sequence alignment or k-mer analysis that are primarily effective in identifying already characterized sequences. Here, we have described an approach that bridges the two extremes. While it is based on sequence-similarity approach, ATLAS provides a similar level of accuracy as phylogenetic approaches while retaining computational efficiency.

ATLAS identifies the ambiguity in the classification of sequences in a sample-specific manner, thereby obviating the need for removing redundancy from the reference database (a computationally expensive process) and ensuring that the method effectively adapts to the specific parameters of the experiment (e.g., choice of hypervariable region in the *16S rRNA* gene). While ATLAS is intended to replace commonly-used “most recent common ancestor” (MRCA) approaches that are unnecessarily conservative, it can also improve on such techniques. The ATLAS partitions are constructed after examining all the query sequences, and after removing spurious connections between database sequences, thereby eliminating many of the errors that can reduce the taxonomic resolution of the MRCA approach.

We have shown that ATLAS is effective in analyzing real microbiome datasets, where it is able to automatically discover taxonomic groupings that are relevant to the interpretation of the data but that do not match predefined taxonomic levels. Examples include subdivisions of the *Bacillus* genus and Clostridial class homology groups. Our paper describes results generated from *16S rRNA* gene sequencing data, however, the approach is applicable to any other marker gene dataset. Because ATLAS relies on marker gene data, it can only provide a level of resolution matching that of the maker gene itself.

Our analysis of the HMP and GEMS datasets reveals a difference in the level of ambiguity identified by ATLAS; our method was able to better resolve the taxonomy of sequences from the HMP project than that of sequences from the GEMS dataset. This finding is likely due to the relationship between the sequences from the two studies and the data found in the reference database. The GEMS study contains data from children from sub-Saharan Africa and Southeast Asia, sequences that are only distantly related to the reference sequences primarily characterized within Western populations. Our findings support the idea that the choice of database plays a huge role in classification accuracy (Nasko et al., 2018). To ensure an accurate taxonomic annotation, a custom environment-specific database is desirable, and the accuracy of the database sequences and their annotation must be ensured. Studies must also carefully consider and document the choice of database.

The GEMS dataset was generated several years ago using 454 sequencing technology with high-insertion-deletion error rates. This can provide useful information for future applications to current long read sequencing datasets, which also have higher insertion-deletion error rates compared to short-read technologies. Despite differences between the GEMS and Bangladesh datasets, ATLAS identified sub-genus partitions in important taxa previously associated with diarrhea. This improved resolution will provide greater insight into potentially harmful or beneficial organisms.

An opportunity for future research is the integration of the approach embodied in ATLAS with phylogenetic algorithms. Phylogenetic approaches can use the partitions identified by ATLAS to prune the reference tree before attempting to place query sequences on the tree, resulting in higher accuracy with lower computational overhead. In the future, we also plan to identify and investigate cases where ATLAS assignments and phylogenetic classifications disagree in order to identify opportunities for improvements to either alignment-based or phylogenetic approaches. As the wealth of microbiome data increases, greater emphasis is being placed on more accurate taxonomic annotations that currently cannot be obtained using fast, sequence similarity-based methods. ATLAS is the first step in this direction.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Human Microbiome Project Data Portal [https://www.hmpdacc.org/HMQCP/] and the GEMS Study of Childhood Diarrheal Disease [http://www.cbcb.umd.edu/datasets/gems-study-diarrheal-disease]. The ATLAS pipeline can be found on GitHub [https://github.com/shahnidhi/outlier_in_BLAST_hits].

AUTHOR CONTRIBUTIONS

NS and MP conceived the research project. NS designed and implemented the algorithm, with the help of JSM and MP. NS and JSM analyzed the data. NS, JSM, and MP wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by the Center for Health-related Informatics and Bioimaging, a Center organized under the MPowering the State Partnership between the University of Maryland Baltimore and College Park campuses. NS, JM, and MP were supported by grants to MP, including grant IIS-1513615 from the NSF.

ACKNOWLEDGMENTS

We thank UMIACS, the Center for Bioinformatics and Computational Biology (CBCB), and the members of the Pop Lab, especially Kassie Wang, for their underlying contributions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01022/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S. F., Wootton, J. C., Zaslavsky, E., and Yu, Y.-K. (2010). The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput. Biol.* 6, e1000852. doi: 10.1371/journal.pcbi.1000852

Barb, J. J., Oler, A. J., Kim, H.-S., Chalmers, N., Wallen, G. R., Cashion, A., et al. (2016). Development of an analysis pipeline characterizing multiple hypervariable regions of 16S rRNA using mock samples. *PLoS One* 11, e0148047. doi: 10.1371/journal.pone.0148047

Bhandari, V., Ahmod, N. Z., Shah, H. N., and Gupta, R. S. (2013). Molecular signatures for *Bacillus* species: demarcation of the *Bacillus* subtilis and *Bacillus* cereus clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. *Int. J. Syst. Evol. Microbiol.* 63, 2712–2726. doi: 10.1099/ijs.0.048488-0

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008. doi: 10.1088/1742-5468/2008/10/P10008

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., and Haussler, D. (1993). Using dirichlet mixture priors to derive hidden markov models for protein families. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1, 47–55.

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

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* 69, 330–339. doi: 10.1016/j.mimet.2007.02.005

Collins, M. D., Lawson, P. A., Willems, A., Cordoba, J. J., Fernandez-Garayzabal, J., Garcia, P., and Farrow, J. A. E. (1994). The phylogeny of the genus Clostridium: proposal of five new genera and eleven new species combinations. *Int. J. Syst. Evol. Microbiol.* 44(4), 812-826.

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

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* 6, e4652. doi: 10.7717/peerj.4652

Fitch, W. M. (1971). Toward Defining the Course of Evolution: minimum change for a specific tree topology. *Syst. Zool.* 20, 406. doi: 10.2307/2412116

Godfray, H. C. J. (2002). Towards taxonomy's "glorious revolution." *Nature* 420, 461. doi: 10.1038/420461a

Johnson, J. L., and Francis, B. S. (1975). Taxonomy of the clostridia: ribosomal ribonucleic acid homologies among the species. *Microbiology*, 88(2), 229-244.

Kieser, S., Sarker, S. A., Sakwinska, O., Foata, F., Sultana, S., Khan, Z., et al. (2018). Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology. *Environ. Microbiol.* 20, 2256–2269. doi: 10.1111/1462-2920.14274

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 12 Suppl 2, S4. doi: 10.1186/1471-2164-12-S2-S4

Liu, Y., Lai, Q., Göker, M., Meier-Kolthoff, J. P., Wang, M., Sun, Y., et al. (2015). Genomic insights into the taxonomic status of the *Bacillus* cereus group. *Sci. Rep.* 5, 14082. doi: 10.1038/srep14082

Lopetuso, L. R., Scaldaferri, F., Petito, V., & Gasbarrini, A. (2013). Commensal Clostridia: leading players in the maintenance of gut homeostasis. Gut pathogens, 5(1), 23.

Nasko, D. J., Koren, S., Phillippy, A. M., and Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* 19:165. doi: 10.1186/s13059-018-1554-6

Nguyen, N.-P., Mirarab, S., Liu, B., Pop, M., and Warnow, T. (2014). TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* 30, 3548–3555. doi: 10.1093/bioinformatics/btu721

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16, 236. doi: 10.1186/s12864-015-1419-2

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

Pop, M., Walker, A. W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M. A., et al. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol.* 15, R76. doi: 10.1186/gb-2014-15-6-r76

Rasko, D. A., Altherr, M. R., Han, C. S., and Ravel, J. (2005). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol. Rev.* 29, 303–329. doi: 10.1016/j.femsre.2004.12.005

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

Shah, N., Altschul, S. F., and Pop, M. (2018). Outlier detection in BLAST hits. *Algorithms Mol. Biol.* 13, 7. doi: 10.1186/s13015-018-0126-3

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Walker, A. W., Duncan, S. H., Louis, P., and Flint, H. J. (2014). Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol.* 22, 267–274. doi: 10.1016/j.tim.2014.03.001

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

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

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* 13, 303–314. doi: 10.1038/nrg3186

Zhang, W., and Sun, Z. (2008). Random local neighbor joining: a new method for reconstructing phylogenetic trees. *Mol. Phylogenet. Evol.* 47, 117–128. doi: 10.1016/j.ympev.2008.01.019

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

frontiers
in Genetics

ORIGINAL RESEARCH
published: 17 October 2019
doi: 10.3389/fgene.2019.00989

Metabolic Overlap in Environmentally Diverse Microbial Communities

Eric R. Hester, Mike S. M. Jetten, Cornelia U. Welte and Sebastian Lücker*

Department of Microbiology, Radboud University, Nijmegen, Netherlands

The majority of microbial communities consist of hundreds to thousands of species, creating a massive network of organisms competing for available resources within an ecosystem. In natural microbial communities, it appears that many microbial species have highly redundant metabolisms and seemingly are capable of utilizing the same substrates. This is paradoxical, as theory indicates that species requiring a common resource should outcompete one another. To better understand why microbial species can coexist, we developed metabolic overlap (MO) as a new metric to survey the functional redundancy of microbial communities at the genome scale across a wide variety of ecosystems. Using metagenome-assembled genomes, we surveyed nearly 1,000 studies across nine ecosystem types. We found the highest MO in extreme (i.e., low pH/high temperature) and aquatic environments, while the lowest MO was observed in communities associated with animal hosts, the built/engineered environment, and soil. In addition, different metabolism subcategories were explored for their degree of MO. For instance, overlap in nitrogen metabolism was among the lowest in animal and engineered ecosystems, while species from the built environment had the highest overlap. Together, we present a metric that utilizes whole genome information to explore overlapping niches of microbes. This provides a detailed picture of potential metabolic competition and cooperation between species present in an ecosystem, indicates the main substrate types sustaining the community, and serves as a valuable tool to generate hypotheses for future research.

Keywords: bioinformatics, metagenomics, microbial communities, metagenome assembled genomes (MAGs), niche, functional redundancy

OPEN ACCESS

Edited by:
Joao Carlos Setubal,
University of São Paulo, Brazil

Reviewed by:
Anna Heintz-Buschart,
German Center for Integrative Biodiversity Research, Germany
Christoph Kaleta,
University of Kiel, Germany

****Correspondence:***
Eric R. Hester
ericokh@gmail.com

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

Received: *06 June 2019*
Accepted: *17 September 2019*
Published: *17 October 2019*

Citation:
Hester ER, Jetten MSM, Welte CU and Lücker S (2019) Metabolic Overlap in Environmentally Diverse Microbial Communities. Front. Genet. 10:989. doi: 10.3389/fgene.2019.00989

INTRODUCTION

Microorganisms drive global biogeochemical cycles, but they do not work or live in isolation. In order for any living species to survive, they must engage in competition for space and resources with other organisms that share similar nutritional requirements. The concept of loss of species less adapted relative to their competitors is known as competitive exclusion (Gause, 1934). When one species cannot sufficiently persist in a habitat, they become locally extinct. Through selection of traits that reduce the dependence on a common resource, populations may shift toward coexistence. This is known as niche partitioning, whereby competition is avoided through the utilization of different resources (Schoener, 1974). Evidence that these ecological and evolutionary forces shape microbial communities is prevalent in literature; however, the strength of these forces varies with the availability of resources [reviewed in (Nemergut et al., 2013)].

Describing a niche of an organism has remained challenging ever since the concept first emerged (Hutchinson, 1957). Typically, closely related species are thought to share similar niches, assuming their evolutionary relatedness is reflected in their nutritional requirements (Langille et al., 2013). Recently, neutral genetic markers have emerged as a proxy to measure species' divergence on an evolutionary timescale; however, these phylogenetic markers (i.e., 16S rRNA genes) are unsuitable to evaluate differences in the biochemical capacity of the organisms (Caro-Quintero and Konstantinidis, 2012). Whole genomes contain information relevant to the metabolic capacity of a species, which is essential to describe the putative niches a microbial species may occupy. If one were to ask about the overlap of two microorganisms' niches, it is conceivable that this is akin to asking how similar the two are on a genomic level. Consequently, the metabolic niche of an organism can be predicted from the genome. However, the metabolic niche must be distinguished from the fundamental niche, which includes factors such as morphological features or transcriptional and translational regulation. These features also influence an organism's adaptation and persistence in a community, but their inclusion introduces additional complexities that are largely absent from genomics-based investigations.

With the continued advancement in high-throughput DNA sequencing, large amounts of genomic data are frequently released and available for public use. Several recent publications have reported thousands of novel bacterial and archaeal metagenome-assembled genomes (MAGs; Anantharaman et al., 2016; Parks et al., 2017; Delmont et al., 2018; Tully et al., 2018). The sequencing data originated from hundreds of studies investigating different ecosystems, such that these genomes represent a diverse set of taxa from ecosystems around the globe. This presents an opportunity to address the following important questions: how variable is niche overlap in microbial communities across different ecosystems and does the nature of the overlap (i.e., abundance of genes involved in nitrogen cycling) change based on habitat?

In the current study, we surveyed niche overlap in microbial communities by searching for shared pathways in the metabolic reaction network of species within these communities, which we refer to as "metabolic overlap" (MO). This approach was used to investigate two main questions. First, does the degree of niche overlap in microbial communities vary between ecosystems (i.e., do some communities have more species that utilize the same substrates)? Second, how do these microbial communities vary in the degree of overlap of different metabolic categories (e.g., nitrogen or sulfur metabolism)?

We observed patterns of overlap in microbial community members' metabolism across different ecosystems, which were largely consistent with literature reports (Martiny et al., 2006; Kelly et al., 2014; Reese et al., 2018). For instance, a low degree of MO was found in microorganisms involved in highly specialized animal host–microbe associations, while aquatic microbes displayed a cosmopolitan repertoire of strategies for nutrient acquisition. These variations seem to be driven by different categories of metabolism, depending on the ecosystem. In addition, we addressed the question of how much the phylogenetic relationship of microbes corresponds to their MO. We found that phylogenetic distance between microorganisms was indeed a good predictor for the degree of MO. The strength of this relationship, however, varied between different ecosystems. Generally, survey-based metrics like MO enable observations of global trends and prompt fundamental questions about the biology and ecology of microorganisms.

RESULTS

Definition of MO

We defined MO as the number of compounds (i.e., reactants) that can be utilized by two organisms based on their shared metabolic network (**Figure 1**). For example, an organism (Org_1) that can perform all steps of denitrification from nitrate (NO_3^-) to nitrogen gas (N_2, four reactions in total) shares two reactants with a partially denitrifying organism (Org_2) that only reduces NO_2^- to N_2O. This then results in a MO = 2 (ignoring the rest of their metabolism). To obtain a value that reflects the degree in which species in a community have overlapping niches, we calculated the median MO between all MAGs in a given study. These studies were grouped into distinct ecosystems based on their origin (**Figure 2**, **Table 1**). Conceivably, identifying MO allows a broad identification of species with overlapping niches by counting the compounds that link complementary

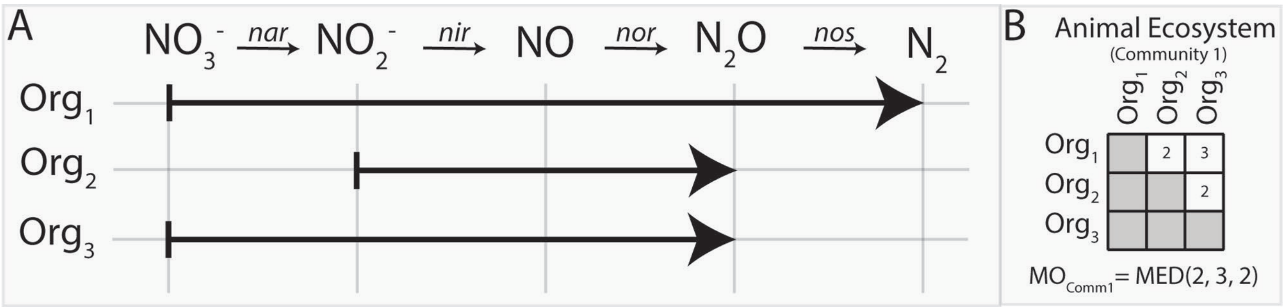

A
$NO_3^- \xrightarrow{nar} NO_2^- \xrightarrow{nir} NO \xrightarrow{nor} N_2O \xrightarrow{nos} N_2$
Org_1
Org_2
Org_3
B
Animal Ecosystem
(Community 1)
Org_1 Org_2 Org_3
2
3
2
MO_{Comm1} = MED(2, 3, 2)

FIGURE 1 | Metabolic overlap is a metric that compares the overlap in the metabolism of two organisms by calculating the number of reactants these species can utilize in common. This is determined by establishing their shared biochemical pathways **(A)** and counting which reactants both can use in common (i.e., common reactants utilized by organisms 1 and 2 is NO_2^- and NO; thus, the $MO_{org(1,2)}$ = 2). The number of substrates shared between a set of organisms is represented in a matrix **(B)**. Once all pairwise MO comparisons have been made for a community, the median metabolic overlap can be calculated.

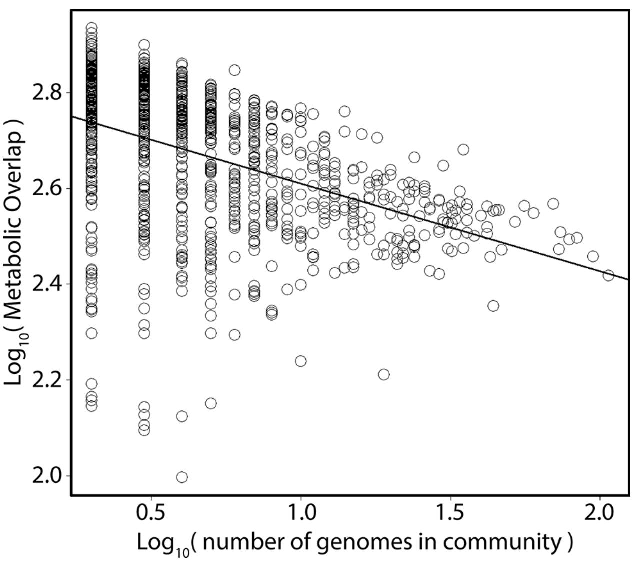


FIGURE 2 | Relationship between metabolic overlap and the number of genomes in a community. Each point represents one of the 962 studies. The *x* axis depicts the total number of MAGs in a given study; the *y* axis, the mean metabolic overlap of that study.

metabolic pathways. As the metabolic routes used to degrade certain substrates can vary between organisms, counting the number of shared reactants will reveal MOs that would not be uncovered by shared reactions only. Furthermore, as the number of reactants can vary between reactions, this approach is more sensitive in identifying weak metabolic similarities between organisms.

We acknowledge that previous efforts to predict microbe–microbe interactions within microbial communities have been made with similar logic to the current approach. In particular, the NetCooperate software, utilizing the NetSeed framework, is a method to identify putative interactions in a community. It does so by using genome information to predict auxotrophies of the organisms present, based on the incompleteness of certain biosynthesis pathways leading to a dependency of the respective organism to external sources of the lacking metabolite (Carr and Borenstein, 2012; Levy et al., 2015). Thus, the NetSeed/NetCooperate approach predicts complementarity between species, which consequently occupy distinct niches, while the goal of our MO approach is to identify to what extent two species fill a common niche.

Metabolic Overlap of Microbial Communities in Different Ecosystems

In order to survey the degree of MO in various ecosystems from around the globe, thereby identifying the degree in which microbial species within the community overlap in the niches they fill, the set of Uncultivated Bacteria and Archaea (UBA) MAGs published by Parks et al. (2017) was utilized. Contrasting to the naming scheme, this set contained some MAGs of cultured species also. The average predicted genome completeness of these MAGs ranged from 50% to 100%. A completion-based inclusion threshold of MAGs was found to have a negligible impact on the average MO of communities (**Supplemental Figure 1**). In contrast, the number of MAGs included drastically decreased as a result of a more stringent threshold on genome completeness, resulting in ecosystems poorly or not at all represented (**Supplemental Figure 2**). Several studies included in the UBA dataset included only one MAG and were excluded from our analyses. In total, 6,727 MAGs from the Parks et al. dataset, representing 962 studies, were included (**Table 1**). Studies were classified into their respective ecosystems of origin based on information included in the submission to the public repository or by manual curation if this information was insufficient. This resulted in nine ecosystem categories (**Table 1**). In total, the reaction space consisted of 1,386 unique compounds predicted to be utilized by the organisms represented by the current set of MAGs.

In a given ecosystem, MO and the predicted average genome sizes of MAGs were strongly correlated (**Supplemental Figure 3**; $p < 0.01$). In addition, average genome sizes significantly varied between ecosystems (**Supplemental Figure 4**; ANOVA; $F = 88$; $p < 0.001$). The average predicted genome sizes were the highest in studies from the built environment (4 ± 0.65 Mbp) and lowest in extreme environments (2 ± 0.96 Mbp). The number of MAGs in a given community (grouped per study) negatively correlated with the average MO of the community (**Figure 2**; Kendall $\tau = -0.38$; $p < 0.001$). As we were interested in investigating how

TABLE 1 | Number of studies and metagenomes within each ecosystem.

	Fresh water	Brackish	Extreme	Marine	Built	Animal	Engineered	Plant	Soil
Amino Acid	4.97E-05	4.24E-05	5.22E-05	4.63E-05	4.06E-05	3.33E-05	3.59E-05	4.09E-05	3.52E-05
Aromatic	6.61E-06	3.26E-06	6.55E-06	8.59E-06	6.91E-06	1.05E-06	2.80E-06	4.43E-06	6.02E-06
carbohydrates	5.58E-05	5.29E-05	6.03E-05	5.42E-05	5.09E-05	4.64E-05	4.33E-05	4.53E-05	4.31E-05
Cofactors	5.27E-05	4.71E-05	4.99E-05	4.79E-05	4.15E-05	3.20E-05	3.34E-05	4.12E-05	3.68E-05
Fatty acids	6.49E-05	7.01E-05	6.32E-05	6.13E-05	5.58E-05	5.33E-05	4.67E-05	5.07E-05	4.35E-05
Nitrogen	4.80E-06	4.90E-06	4.17E-06	3.71E-06	4.40E-06	2.02E-06	2.40E-06	2.63E-06	3.37E-06
Nuleoside	2.27E-05	1.82E-05	2.39E-05	2.28E-05	1.97E-05	2.46E-05	2.29E-05	1.86E-05	1.89E-05
Nuelotide sugars	5.01E-06	4.57E-06	5.38E-06	4.01E-06	3.78E-06	4.62E-06	4.51E-06	3.10E-06	4.43E-06
Phosphorus	4.62E-06	4.07E-06	2.91E-06	3.87E-06	3.50E-06	3.58E-06	3.24E-06	1.93E-06	3.05E-06
Protein	1.88E-05	1.75E-05	2.50E-05	1.62E-05	1.29E-05	1.82E-05	1.63E-05	1.66E-05	1.49E-05
Respiration	8.11E-06	8.48E-06	7.24E-06	7.12E-06	6.27E-06	2.98E-06	4.87E-06	4.74E-06	5.34E-06
Secondary Metabolism	2.40E-06	2.11E-06	3.93E-06	2.29E-06	1.80E-06	1.88E-06	1.95E-06	3.35E-06	2.28E-06
Sulfur	3.10E-06	2.88E-06	2.65E-06	3.26E-06	4.17E-06	9.78E-07	1.45E-06	9.84E-07	2.34E-06

MO varied between ecosystems, irrespective of the differences in genome sizes between ecosystems, we normalized MO to the median genome size of the respective study (**Figure 3**). MO was found to vary significantly between ecosystems (χ^2 = 75.3; $p < 0.001$). Communities from animal, built, engineered, and soil ecosystems had significantly lower MO than aquatic ecosystems ($p < 0.05$; **Figure 3**, **Supplemental Table 1**). Furthermore, extreme ecosystems had significantly higher MO than built and engineered ecosystems ($p < 0.05$; **Figure 3**, **Supplemental Table 1**).

Breakdown of MO Scores Across Different Ecosystems to Different Levels of Metabolism

To investigate how MO varied between ecosystems within different categories of metabolism (SEED subsystems), the MO within these subcategories was determined for each ecosystem and compared to the average value of all ecosystems (**Supplemental Table 2**). All metabolic subsystems varied between ecosystems (Kruskal–Wallis; $p < 0.001$; **Supplemental Table 2**). Animal, built, and engineered ecosystems in general had a lower MO for the majority of subcategories of metabolism with a few exceptions (Dunn; $p < 0.01$; **Supplemental Tables 3–15**). In contrast, communities from the engineered ecosystems had higher MO in protein and nucleotide sugar metabolism, as did communities from animal ecosystems. While most subcategories of metabolism from the built environment had lower MO than other ecosystems, these communities contained higher MO in nitrogen and sulfur metabolism (**Supplemental Table 16**). In contrast to the above communities, which were dominated by lower than average MO scores, extreme, freshwater, and marine ecosystems had higher than average MO scores in the majority of the categories of metabolism (**Supplemental Tables 3–15**).

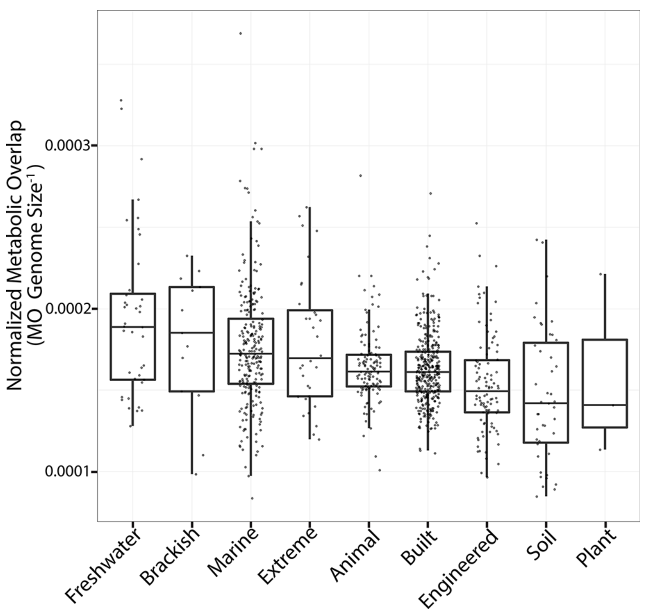


FIGURE 3 | Metabolic overlap across all ecosystems. Boxplots are plotted with the black bar representing the median, the box corresponds to the 25% and 75% quartiles, and the whiskers are the extreme values. The *y* axis is MO normalized by genome size to account for differences between median genome sizes across ecosystems. The ecosystems are sorted from left to right based on the median MO. Each point represents the median MO of all MAGs from a given study.

Nitrogen metabolism was used to further investigate the influence of partial pathways on the MO. Therefore, the ratios of complete to partial denitrifiers were calculated for all ecosystems (i.e., complete denitrifiers encoding all proteins required for NO_3^-, NO_2^-, NO, and N_2O reduction; partial denitrifiers missing at least one gene; **Figure 4A**). The proportion of MAGs containing at least one denitrification gene ranged between ecosystems, with the lowest in the animal ecosystem and the highest in the built environment (**Figure 4B**). The built environment contained one of the highest MO in nitrogen metabolism and also had the highest ratio of complete to partial denitrifiers of all other ecosystems (**Figure 4C**). Contrary, the animal ecosystem, which by far had the lowest MO in this category, also contained mostly partial denitrifiers (**Figure 4C**).

Phylogenetic Relationship of Organisms and Its Relationship to the MO

In order to determine if the evolutionary relatedness between MAGs was correlated with MO, the UBCG pipeline was utilized to infer a phylogenetic tree based on a concatenated alignment of 92 universal bacterial marker genes (Na et al., 2018). A significant

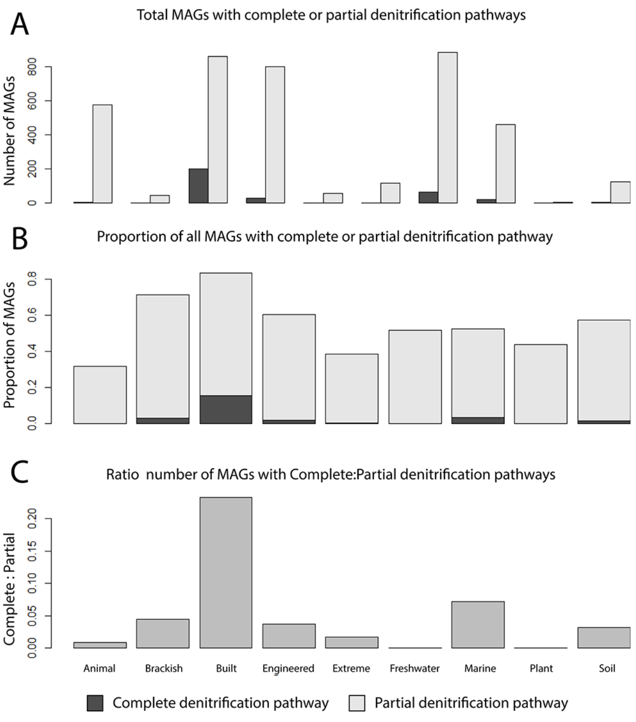


FIGURE 4 | Proportions of complete and partial denitrifiers across different ecosystems. **(A)** Number of MAGs encoding all proteins to reduce NO_3^- to N_2 (complete denitrifiers) compared to the number of MAGs with one or more of the respective genes missing. **(B)** Proportion of MAGs of the total community that were either partial denitrifiers or complete denitrifiers. **(C)** Ratio of complete to partial denitrification pathways.

negative correlation was observed between phylogenetic distance and MO for all ecosystems (**Figure 5**; *r* = −0.33; $p < 0.001$); however, the strength of this association varied. Phylogenetic distance and MO had the strongest association in plant (*r* = −0.64), built (*r* = −0.53) and marine ecosystems (*r* = −0.47), whereas the lowest associations were seen in animal (*r* = −0.16), extreme (*r* = −0.19) and freshwater ecosystems (*r* = −0.21; **Figure 5**).

DISCUSSION

In the current study, a new metric termed MO, which describes how similar two species' metabolisms are, was developed in the context of a genome-based survey of microbial communities from diverse ecosystems. High MO between two species suggests that they have the capacity to perform similar metabolic reactions and thus have similar growth requirements and fill similar niches. In contrast, low MO suggests that the two species in question may compete for fewer resources. Thus, the average MO of a community can be interpreted such that in a community with high MO many community members are overlapping in their biochemistry and could in theory compete for a similar niche, whereas a low average MO would suggest the opposite.

Ecological and Evolutionary Drivers of MO

There are several well-studied ecological forces that shape microbial community structure. Community diversity is maintained *via* dispersion (immigration and emigration) as well

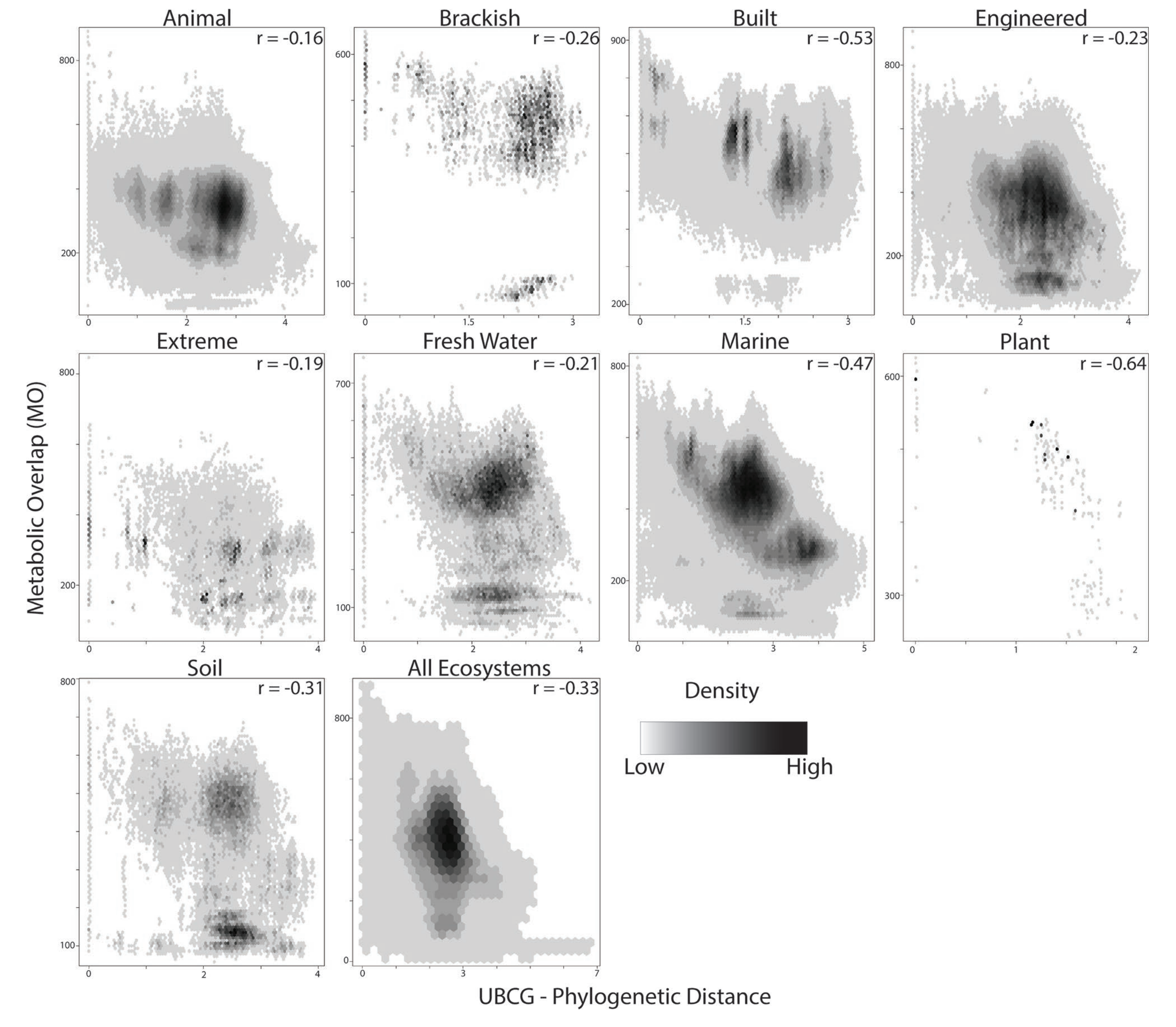


FIGURE 5 | Relationship between metabolic overlap and phylogenetic distance. Each point represents a pairwise comparison between two MAGs. The density of points is represented by a black and white gradient. The Spearman correlation coefficient is indicated in the upper left-hand corner of each plot.

as speciation and extinction. In studying patterns of microbial biogeography, dispersion limitations were seen as one of the driving forces in structuring microbial community patterns in salt marshes and rice paddies and likely have an influence on the genomic adaptations of marine microorganisms (Martiny et al., 2006; Kelly et al., 2014; Lüke et al., 2014). Microbial biogeography theory has also been applied to help understanding compartmentalized host-associated microbial communities such as microbes in the human lungs (Whiteson et al., 2014). In this study, we observed major ecosystem-dependent differences in the MO of microbial community members (**Figure 3**). This variation may in part be attributed to dispersion limitations inherent to each ecosystem, where ecosystems in which the dispersion of microbial community members is limited would have less overlap than open homogenous ecosystems. Accordingly, the highest MO was observed in aquatic ecosystems, namely, communities from the marine open ocean environment, while animal host-associated communities contained some of the lowest MO (**Figure 3**). Ecosystems such as the ocean are likely to not have as strong dispersal limitations as ecosystems like the animal gut or human lungs, and these differences may be a driving force in structuring the MO of their respective microbial communities.

In addition to dispersion as an ecological force, disturbances to ecosystems can also play a large role for species diversity, driving extinction or speciation within the community (Connell, 1978; Buckling et al., 2000). Varying degrees of disruption would impart some signature on the metabolic pathways represented in the microbial community. A higher frequency of disturbance would contribute to the extinction of species and reduce the number of redundant metabolisms in a given system. For example, disturbances associated with the marine ecosystem (high MO) such as storms or temperature anomalies are likely less frequent and intense than the regular consumption of foodstuff or intermittent bouts of inflammation in animal guts (low MO) (Kashyap et al., 2013; David et al., 2014; Reese et al., 2018).

Substrate Spectrum as a Possible Driver of MO in Ecosystems

The availability of resources, both in quality and quantity, drives which species can thrive in a given system. In the open ocean, the input of labile organic matter is a major factor controlling microbial activity in the photic zone, where phototrophs fix large quantities of inorganic carbon, making new organic matter available to heterotrophic organisms (Hansell and Carlson, 2002; Aylward et al., 2015). It is understood that differences in the composition of dissolved organic matter enrich for different clades of microorganisms and that the composition of the community is highly influential on the capacity to degrade this carbon (Nelson et al., 2013; Solden et al., 2018). In the case of animal- and plant-associated microorganisms, the composition of substrates provided to the microorganisms is often host-specific, which is thought to drive species specificity of the microbiota (Berg et al., 2014; Nelson et al., 2013; Hester et al., 2016; Quinlan et al., 2018; Reese et al., 2018; Jones et al., 2019). It would follow that a higher substrate selection would drive diversity in the microbial community, and the higher diversity of substrates would then lead to more diverse microbial metabolisms. In the current study, a negative relationship between the richness of a community (number of genomes in a given sample) and their average MO was observed, which suggests that in more diverse communities there is less MO (**Figure 2**).

In addition to the quality of substrates, the quantity of organic matter also drastically differs between ecosystems. The concentration of dissolved organic carbon (DOC) can vary greatly in aquatic systems, with around 40 µmol L^{-1} DOC in groundwater and 5,000 µmol L^{-1} in swamps and marshes (Søndergaard and Thomas, 2004). Likewise, variations in animal's diet influence the availability of different substrates for microorganisms. In particular, the diet of an animal influences the availability of nitrogen to microbes in animal guts (Reese et al., 2018). Equally, N availability has a strong impact on plant-soil feedbacks, influencing the abundance and metabolism of microorganisms in the rhizosphere (Hester, 2018). If substrates are available in high-enough concentrations, the effect of competition may be reduced, potentially leading to a higher number of species consuming a common substrate (i.e., higher MO). In the current study, we observe microbial communities from animal ecosystems had the lowest overlap in categories of metabolism involved in nitrogen and amino acid metabolism, which corresponds to the idea of N limitations in the animal gut and known auxotrophies (**Supplemental Tables 3 and 8**; Reese et al., 2018; Zengler and Zaramela, 2018). In contrast, microbial communities from the built environment tend to have higher overlap in nitrogen and sulfur metabolism, although the built environment is a loosely defined ecosystem type with limited literature detailing nutrient fluxes through the system (**Supplemental Tables 8 and 15**; Adams et al., 2015). This stark contrast of nitrogen metabolism overlap between the built and animal ecosystems, which both generally displayed a lower than average MO, corresponded to the observed number of species capable of complete denitrification. The built ecosystem had the highest nitrogen metabolism MO, which largely was attributed to the highest proportion of microbial species capable of complete denitrification (**Figure 4**). This was contrasted by the low number of complete denitrifiers in the animal system. While the differences here could be due to nutrient availability, one should also consider possible differences in life strategies for persisting in a particular environment (i.e., detoxification vs. energy conservation).

Influence of Phylogenetic Relationship on MO

Populations that become isolated and diverge on an evolutionary timescale do so as a result of being exposed to different environments and thus different selection pressures on specific traits, although some mechanisms exist that make this divergence less clear (i.e., convergent evolution, horizontal gene transfer, etc.). In the current study, a correlation was observed between the MO of species and their phylogenetic relationship (**Figure 5**), with a reduced MO in taxa that are more distantly related. While this corresponds well to theory, the strength of the relationship between phylogenetic relatedness and MO varied between ecosystems, suggesting that ecological differences between these ecosystems influence this relationship.

The dominant taxonomic groups often vary between different ecosystems as a result of the underlying nutrient profiles or physical properties of those ecosystems. This may be a result of stronger selection pressures in a given ecosystem for traits specific to a few select phylogenetic groups (i.e., methanogenesis, ammonia, and nitrite oxidation), as opposed to traits that are more widespread (i.e., denitrification). Phylogenetic groups may vary in the number of traits (i.e., some groups are more metabolically versatile than others, which often is also reflected in larger genome sizes within these groups), and MO is determined by the number of reactions a given pair of species share. For example, Zimmerman et al. (2013) found that a set of phylogenetically diverse bacteria and Archaea had the potential to produce a subset of three extracellular enzymes. The ability to produce these enzymes was nonrandomly distributed phylogenetically. It follows that ecosystems that have strong selection pressures for metabolically diverse phylogenetic groups would have a weaker relationship between the phylogenetic relatedness and MO. Interestingly, within each ecosystem type, there was a strong positive correlation between genome size and MO (**Supplemental Figure 3**), and the observed negative relationship of phylogenetic distance and MO seemed to be related to genome size (**Figure 5**). The built environment, which contained the largest genomes out of all ecosystems (**Supplemental Figure 4**), also had the strongest negative relationship between phylogenetic distance and MO (**Figure 5**). On the other hand, genomes from the animal ecosystem were the smallest and also showed the weakest relationship between MO and phylogenetic distance. It thus appears that both genome size (i.e., number of genes) and phylogenetic affiliation (closely related species sharing similar pathways) jointly influence MO between a given pair of species.

Caveats and Limitations of Genetic Predictions of MO

The emergence of vast amounts of sequence data has allowed the assembly of genomes of microorganisms from fragmented DNA isolated from the environment. The degree of information in whole genomes compared to that from marker genes (both phylogenetic and metabolic) is likely to provide significant advances in our understanding of the genetic organization of microorganisms. In addition, knowing that a certain set of genomes were physically in the same sample is advantageous in addressing fundamental questions about the ecology and evolution of microbial communities in natural settings. Unfortunately, there are still significant limitations when dealing with MAGs. Specifically, the amount of information lost in the process of genome assembly and binning reduces our understanding of population-level genetic variation. It is still challenging to assemble genomes from organisms of low abundance, in particular when communities are complex (Cleary et al., 2015; Ayling et al., 2019). This narrows our view of genetic linkages between microorganisms toward the highly abundant and thus frequently observed species. However, these are mainly technological limitations, with solutions like long read sequencing becoming more widely available. Additionally, there is a significant lack of information about the environments in which samples were taken in the public archives. For instance, knowing the abundance of an organism in the community would significantly aid in inferring ecological interactions. The absence of such information limits what can be assessed with metrics such as MO and calls for an urgent need to provide as much metadata on samples as possible.

In addition to the technical limitations mentioned above, there are also limitations in methods such as MO, which rely heavily on accurate automated annotation of genetic elements in genomes. Specifically, database quality is a key driver in the accuracy of survey studies such as the one presented here. A major issue is the inability to assign functions to many genes, even in the genomes of the most well-studied microorganisms (35% hypothetical proteins in *Escherichia coli* genome; Ghatak et al., 2019). Apart from the limitations to automatic annotation methods, there are different levels of biology associated with niches that are not captured in genome-level information. These limitations include a lack of information of whether a gene is transcribed, whether the transcript is translated to a functional product, and ultimately variations in affinity and activity of this protein. The variation in transport efficiency and regulatory mechanisms certainly contributes to the competitive advantage of an organism and thus the niche this organism fills. These complexities are not easily derived from genomic information. Complementary techniques, such as transcriptomics, proteomics, and exometabolomics, could supplement the approach presented here by highlighting pathways that are expressed or translated under a given condition. Ideally, as emphasized by Bowers et al. (2017), in order to improve discovery-based approaches that rely on machine readable formats of public repositories, additional information should accompany MAG submissions. This set of information would not only help assess the quality of the genome but aid in associating the genetic information to the biology and ecology of the organism. Ideally, such information should include conditions of the environment from which the species' genome was obtained (i.e., pH and temperature) and, if the species was cultivated, any physiological parameters that may have been measured (i.e., growth rate, substrate usage profile and affinities, etc.).

CONCLUSIONS

The observation of variation in MO across different ecosystems begs several questions about the nature of microbial community metabolism. Specifically, what drives metabolic versatility in microbial communities? Are there generalizable rules that can be deduced? Survey-based studies enriched with additional information, such as those highlighted above, may shed additional light on important factors that drive MO. In addition, there is an urgent need to complement predictions based on the genetics of microorganisms with phenotypic data. Ultimately, understanding drivers of microbial community metabolism will lead to a better ability to predict and engineer microbial communities for industrial or conservational purposes.

METHODS

Data Origin and Annotation of Ecosystems

Metagenome-assembled genomes utilized in the current study comprised the set published by Parks et al. (2017). The UBA MAGs were downloaded from the authors' repository (https://data.ace.uq.edu.au/public/misc_downloads/uba_genomes/). The accompanying data from the UBA MAG set, including CheckM metrics of predicted genome completeness and size, were obtained from the publication (Parks et al., 2017). Each study in the UBA set of MAGs was manually sorted into a set of nine ecosystems.

Metabolic Overlap Calculation

All MAGs were subsequently annotated using a custom pipeline based on the SEED API (Overbeek et al., 2005; Aziz et al., 2008). In brief, protein encoding genes (pegs) were called from the assemblies using svr_call_pegs (http://servers.nmpdr.org/sapling/server.cgi?pod=ServerScripts). Each of these proteins was then assigned to a figfam with svr_assign_using_figfams (our annotations can be found at: ericrhester.com/metabolicOverlap/annotations/results.tar.gz). The association of a protein to a biochemical reaction was then made with svr_roles_to_reactions. Custom script (rxn_expandinfo) associated reactions with compounds from the reaction database, which is found on the ModelSEED git repository (https://github.com/ModelSEED). Finally, the number of compounds shared between two organisms' set of biochemical reactions is calculated to create a pairwise MO score, and an overlap matrix was constructed to store this information. This was made using the custom python scripts rxn_to_connections and lists_to_matrix, respectively (https://github.com/ericHester/metabolicOverlap). The overlap matrix represents the MO of all organisms within a single community and the average MO of all of these organisms is utilized in comparison in this study.

In addition to an overall MO score for a community, the above approach was used to calculate the MO of various subcategories of metabolism for the respective community. In addition to the above, an additional step was performed where pegs were assigned to their respective SEED subsystems and filtered with a custom script utilizing svr_roles_to_subsys. With pegs assigned to these metabolic categories, the above pipeline was used to identify reactions and compounds shared between pairs of organisms, subsequently resulting in an overlap matrix similar to that above. In this case, the overlap matrix stores the MO of the community pertaining to a specific category of metabolism. Matrices and accompanying data were further analyzed in R (R Core Team, 2016).

Relating Phylogenetic Distances of Mags to Their MO Within Communities

In order to associate the phylogenetic distance of assembled genomes to their MO, the UBCG pipeline was utilized (Na et al., 2018). This pipeline extracts 92 conserved phylogenetic marker genes and builds multiple alignments for each gene. The resulting alignments are concatenated, and a maximum likelihood tree is inferred. This tree was imported into R utilizing the *ape* package, and distances were extracted from the tree object with the *cophenetic* function (Paradis et al., 2004). The result is a distance matrix containing phylogenetic distances between each pair of MAGs. Subsequently, this phylogenetic distance matrix and the overlap matrix storing MO scores were correlated using the *mantel.test* function from the ape package. The Spearman rank correlation coefficient was calculated for each ecosystem subset.

DATA AVAILABILITY STATEMENT

The Uncultured Bacterial and Archaeal (UBA) MAGs were downloaded from the author's repository (https://data.ace.uq.edu.au/public/misc_downloads/uba_genomes/). The accompanying data from the UBA MAG set, including CheckM metrics of predicted genome completeness and size, was obtained from the publication (Parks et al., 2017).

AUTHOR CONTRIBUTIONS

EH, SL, CW, and MJ designed the study. EH performed the analysis and drafted the manuscript. EH, SL, CW, and MJ contributed to the editing of the manuscript.

ACKNOWLEDGMENTS

We would like to acknowledge Michiel van den Heuvel for contributing to the metabolic overlap pipeline. Funding was provided by the European Research Council (ERC AG ecomom 339880) and the Netherlands Organisation for Scientific Research (NWO; Gravitation Grant SIAM 024.002.002, VIDI grant 016.Vidi.189.050)

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00989/full#supplementary-material

SUPPLEMENTAL FIGURE 1 | Relationship between the genome completeness and the average metabolic overlap observed (colored lines, right axis). Each point represents the median MO of a study, which includes MAGs with a given genome completeness or higher.

SUPPLEMENTAL FIGURE 2 | The number of MAGs included in the analysis at a given genome completeness cutoff.

SUPPLEMENTAL FIGURE 3 | Relationship between metabolic overlap and genome size. Each point represents one study. The *y* axis indicates the median metabolic overlap of all MAGs in one study, and the median genome size for all MAGs in this study is on the *x* axis.

SUPPLEMENTAL FIGURE 4 | Genome sizes across ecosystems. The black bar of the boxplot indicates the median, the box edge represents the upper and lower quartiles, whiskers denote extreme values, and individual points are outliers.

SUPPLEMENTAL TABLE 1 | Multiple comparisons for differences in median MO for each ecosystem.

SUPPLEMENTAL TABLE 2 | Kruskal–Wallis test statistics for differences in metabolic overlap for different categories of metabolisms grouped by SEED subsystems.

SUPPLEMENTAL TABLE 3 | Dunn multiple comparisons evaluating the amino acid metabolism subsystem.

SUPPLEMENTAL TABLE 4 | Dunn multiple comparisons evaluating the aromatic compound metabolism subsystem.

SUPPLEMENTAL TABLE 5 | Dunn multiple comparisons evaluating the carbohydrate metabolism subsystem.

SUPPLEMENTAL TABLE 6 | Dunn multiple comparisons evaluating the cofactors metabolism subsystem.

SUPPLEMENTAL TABLE 7 | Dunn multiple comparisons evaluating the fatty acids metabolism subsystem.

SUPPLEMENTAL TABLE 8 | Dunn multiple comparisons evaluating the nitrogen metabolism subsystem.

SUPPLEMENTAL TABLE 9.| Dunn multiple comparisons evaluating the nucleoside metabolism subsystem.

SUPPLEMENTAL TABLE 10 | Dunn multiple comparisons evaluating the nucleotide sugar metabolism subsystem.

SUPPLEMENTAL TABLE 11 | Dunn multiple comparisons evaluating the phosphorus metabolism subsystem.

SUPPLEMENTAL TABLE 12 | Dunn multiple comparisons evaluating the protein metabolism subsystem.

SUPPLEMENTAL TABLE 13 | Dunn multiple comparisons evaluating the respiration subsystem.

SUPPLEMENTAL TABLE 14 | Dunn multiple comparisons evaluating the secondary metabolism subsystem.

SUPPLEMENTAL TABLE 15 | Dunn multiple comparisons evaluating the sulfur metabolism subsystem.

SUPPLEMENTAL TABLE 16 | Summary of the median metabolic overlap for each subsystem of metabolism for all ecosystems.

REFERENCES

Adams, R. I., Bateman, A. C., Bik, H. M., and Meadow, J. F. (2015). Microbiota of the indoor environment: a meta-analysis. *Microbiome* 3 (1), 49. doi: 10.1186/s40168-015-0108-3

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7, 13219. doi: 10.1038/ncomms13219

Ayling, M., Clark, M. D., and Leggett, R. M. (2019). New approaches for metagenome assembly with short reads. *Brief. Bioinformat.* 1–11. doi: 10.1093/bib/bbz020

Aylward, F. O., Eppley, J. M., Smith, J. M., Chavez, F. P., Scholin, C. A., and DeLong, E. F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl. Acad. Sci. U. S. A.* 112 (17), 5443–5448. doi: 10.1073/pnas.1502883112

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 (1), 75. doi: 10.1186/1471-2164-9-75

Berg, G., Grube, M., Schloter, M., and Smalla, K. (2014). The plant microbiome and its importance for plant and human health. *Front. Microbiol.* 5, 1. doi: 10.3389/fmicb.2014.00491

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and Archaea. *Nat. Biotechnol.* 35 (8), 725–731. doi: 10.1038/nbt.3893

Buckling, A., Kassen, R., Bell, G., and Rainey, P. B. (2000). Disturbance and diversity in experimental microcosms. *Nature* 408 (6815), 961–964. doi: 10.1038/35050080

Caro-Quintero, A., and Konstantinidis, K. T. (2012). Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.* 14 (2), 347–355. doi: 10.1111/j.1462-2920.2011.02668.x

Carr, R., and Borenstein, E. (2012). NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. *Bioinformatics (Oxford, England)* 28 (5), 734–735. doi: 10.1093/bioinformatics/btr721

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., et al. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33 (10), 1053–1060. doi: 10.1038/nbt.3329

Connell, J. H. (1978). Diversity in tropical rain forests and coral reefs. *Science*. 199 (4335), 1302–1310. doi: 10.1126/science.199.4335.1302

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 (7484), 559–563. doi: 10.1038/nature12820

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. TM., Rappé, M. S., et al. (2018). Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* 3 (7), 804–813. doi: 10.1038/s41564-018-0176-9

Gause, G. F. (1934). *The struggle for existence*. Gause. GF, editor. . Baltimore: The Williams & Wilkins company. http://www.biodiversitylibrary.org/bibliography/4489 (August 2, 2018). doi: 10.5962/bhl.title.4489

Ghatak, S., King, Z. A., Sastry, A., and Palsson, B. O. (2019). The Y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res.* 47 (5), 2446–2454. doi: 10.1093/nar/gkz030

Hansell, D. A., and Carlson, C. A. (2002). *Biogeochemistry of marine dissolved organic matter*. Elsevier: Academic Press. http://agris.fao.org/agris-search/search.do?recordID = US201300072786 (December 1, 2018).

Hester, E. R., Barott, K. L., Nulton, J., Vermeij M. J. A., and Rohwer, F. L. (2016). Stable and sporadic symbiotic communities of coral and algal holobionts. *ISME Journal* 10 (5), 1157–1169. doi: 10.1038/ismej.2015.190

Hester, E. R., Harpenslager, S. F., van Diggelen J. M. H., Lamers L. L., Jetten, M. S. M., Lüke, C., et al. (2018). Linking nitrogen load to the structure and function of wetland soil and rhizosphere microbial communities. *mSystems* 3 (1), e00214–e00217. doi 10.1128/mSystems.00214-17

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harb Symp. Quant. Biol.* 22 (0), 415–427. doi: 10.1101/SQB.1957.022.01.039

Jones, P., Garcia, B. J., Furches A., Tuskan, G. A., and Jacobson, D. (2019). Plant host-associated mechanisms for microbial selection. *Front. Plant Sci.* 10, 862. doi: 10.3389/fpls.2019.00862

Kashyap, P. C., Marcobal, A., Ursell, L. K., Larauche, M., Duboc, H., Earle, K. A., et al. (2013). Complex interactions among diet, gastrointestinal transit, and gut microbiota in humanized mice. *Gastroenterology* 144 (5), 967–977. doi: 10.1053/j.gastro.2013.01.047

Kelly, L. W. Williams, G. J., Barott, K. L., Carlson, C. A., Dinsdale, E. A., Edwards, R. A., et al. (2014). Local genomic adaptation of coral reef-associated microbiomes to gradients of natural variability and anthropogenic stressors. *Proc. Natl. Acad. Sci. U. S. A.* 111 (28), 10227–10232. doi: 10.1073/pnas.1403319111

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 (9), 814–821. doi: 10.1038/nbt.2676

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 (1), 164. doi: 10.1186/s12859-015-0588-y

Lüke, C., Frenzel P., Ho, A., Fiantis, D., Schad, P., Schneider, B., et al. (2014). Macroecology of methane-oxidizing bacteria: the β-diversity of PmoA genotypes in tropical and subtropical rice paddies. *Environ. Microbiol.* 16 (1), 72–83. doi: 10.1111/1462-2920.12190

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4 (2), 102–112. doi: 10.1038/nrmicro1341

Na, S.-I., Ouk, Y., Seok-Hwan K., Sung-min, Y., Inwoo, H., Chun, B. J. (2018). UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56 (4), 280–285. doi: 10.1007/s12275-018-8014-6

Nelson, C. E., Goldberg, S. J., Kelly, L. W., Haas, A. F., Smith, J. E., Rohwer, F., et al. (2013). Coral and macroalgal exudates vary in neutral sugar composition and differentially enrich reef bacterioplankton lineages. *ISME Journal* 7 (5), 962–979. doi: 10.1038/ismej.2012.161

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., et al. (2013). Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev* 77 (3), 342–356. doi: 10.1128/MMBR.00051-12

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33 (17), 5691–5702. doi: 10.1093/nar/gki866

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20 (2), 289–290. doi: 10.1093/bioinformatics/btg412

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2 (11), 1533–1542. doi: 10.1038/s41564-017-0012-7

Quinlan, Z. A., Remple, K., M. Fox, D., Silbiger, N, J., Oliver, T. A., Putnam, H. M., et al. (2018). Fluorescent organic exudates of corals and algae in tropical reefs are compositionally distinct and increase with nutrient enrichment. *Limnol. Oceanogr. Lett.* 3 (4), 331–340. doi: 10.1002/lol2.10074

R Core Team (2016). *"R: a language and environment for statistical computing."* Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Reese, A. T., Pereira, F. C., Schintlmeister, A., Berry, D., Wagner, M., Hale, L. P., et al. (2018). Microbial nitrogen limitation in the mammalian large intestine. *Nat. Microbiol.* 3 (12), 1441–1450. doi: 10.1038/s41564-018-0267-7

Schoener, T. W. (1974). Resource partitioning in ecological communities. *Science* 185 (4145), 27–39. doi: 10.1126/science.185.4145.27

Solden, L. M., Naas, A. E., Roux, S., Daly, R. A., Collins, W. B., Nicora, C. D., et al. (2018). Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* 3 (11), 1274–1284. doi: 10.1038/s41564-018-0225-4

Søndergaard, M., and Thomas, D. N. (2004). "Dissolved organic matter (DOM) in aquatic ecosystems," in *A study of European catchments and coastal waters, The DOMAINE project*. Eur. Union Domain Proj. 76, 1–76.

Tully, B. J., Graham, E. D., and Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* 5, 170203. doi: 10.1038/sdata.2017.203

Whiteson, K. L., Bailey, B., Bergkessel M., Conrad, D., Delhaes, L., Felts, B., et al. (2014). The upper respiratory tract as a microbial source for pulmonary infections in cystic fibrosis. Parallels from island biogeography. *Am. J. Respir. Crit. Care Med.* 189 (11), 1309–1315. doi: 10.1164/rccm.201312-2129PP

Zengler, K., and Zaramela, L. S. (2018). The social network of microorganisms — how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* 16 (6), 383–390. doi: 10.1038/s41579-018-0004-5

Zimmerman, A. E., Martiny, A. C., and Allison, S. D. (2013). Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. *ISME Journal* 7 (6), 1187–1199. doi: 10.1038/ismej.2012.176

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

frontiers
in Genetics

REVIEW
published: 17 January 2020
doi: 10.3389/fgene.2019.01344

Integrating Computational Methods to Investigate the Macroecology of Microbiomes

Rilquer Mascarenhas[1], *Flávia M. Ruziska*[1], *Eduardo Freitas Moreira*[1], *Amanda B. Campos*[1], *Miguel Loiola*[1], *Kaike Reis*[2], *Amaro E. Trindade-Silva*[1,3], *Felipe A. S. Barbosa*[1], *Lucas Salles*[4], *Rafael Menezes*[3,5], *Rafael Veiga*[6], *Felipe H. Coutinho*[7], *Bas E. Dutilh*[8,9], *Paulo R. Guimarães Jr.*[10], *Ana Paula A. Assis*[10], *Anderson Ara*[11], *José G. V. Miranda*[5], *Roberto F. S. Andrade*[5,6], *Bruno Vilela*[1] and *Pedro Milet Meirelles*[1,3]*

[1] Institute of Biology, Federal University of Bahia, Salvador, Brazil, [2] Chemical Engineering Department, Polytechnic School of Federal University of Bahia, Salvador, Brazil, [3] Department of Ecology, Biosciences Institute, University of Sao Paulo, Sao Paulo, Brazil, [4] Institute of Geology, Federal University of Bahia, Salvador, Brazil, [5] Institute of Physics, Federal University of Bahia, Salvador, Brazil, [6] Center of Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Muniz, Fundação Oswaldo Cruz, Brazil, [7] Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández de Elche, San Juan de Alicante, Spain, [8] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands, [9] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, Netherlands, [10] Department of Ecology, Biosciences Institute, University of Sao Paulo, Butantã, Brazil, [11] Institute of Mathematics, Federal University of Bahia, Salvador, Brazil

OPEN ACCESS

Edited by:
Yi Zhao,
Beijing University of Chinese Medicine, China

Reviewed by:
Ramona Lynn Walls,
University of Arizona, United States
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

****Correspondence:***
Pedro Milet Meirelles
pmeirelles@ufba.br

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

Received: *31 December 2018*
Accepted: *09 December 2019*
Published: *17 January 2020*

Citation:
Mascarenhas R, Ruziska FM, Moreira EF, Campos AB, Loiola M, Reis K, Trindade-Silva AE, Barbosa FAS, Salles L, Menezes R, Veiga R, Coutinho FH, Dutilh BE, Guimarães PR Jr., Assis APA, Ara A, Miranda JGV, Andrade RFS, Vilela B and Meirelles PM (2020) Integrating Computational Methods to Investigate the Macroecology of Microbiomes. Front. Genet. 10:1344. doi: 10.3389/fgene.2019.01344

Studies in microbiology have long been mostly restricted to small spatial scales. However, recent technological advances, such as new sequencing methodologies, have ushered an era of large-scale sequencing of environmental DNA data from multiple biomes worldwide. These global datasets can now be used to explore long standing questions of microbial ecology. New methodological approaches and concepts are being developed to study such large-scale patterns in microbial communities, resulting in new perspectives that represent a significant advances for both microbiology and macroecology. Here, we identify and review important conceptual, computational, and methodological challenges and opportunities in microbial macroecology. Specifically, we discuss the challenges of handling and analyzing large amounts of microbiome data to understand taxa distribution and co-occurrence patterns. We also discuss approaches for modeling microbial communities based on environmental data, including information on biological interactions to make full use of available Big Data. Finally, we summarize the methods presented in a general approach aimed to aid microbiologists in addressing fundamental questions in microbial macroecology, including classical propositions (such as "everything is everywhere, but the environment selects") as well as applied ecological problems, such as those posed by human induced global environmental changes.

Keywords: microbial community modeling, microbial macroecology, spatial scales, machine learning, co-occurrence networks

INTRODUCTION

The purpose of macroecology is to describe spatial patterns of species distribution and abundance, as well as the mechanisms underlying such patterns (McGill, 2003; McGill and Nekola, 2010). The availability of large amounts of data (Hampton et al., 2013) has helped to uncover global ecological patterns in species distribution and abundance, greatly advancing the field of macroecology. This is highlighted by several studies discussing the contribution of microbial community investigations to a unified macroecological theory (Barberán et al., 2014; Blaser et al., 2016; Nelson et al., 2016; Shade et al., 2018). Strong evidence suggests that micro-organisms in deep display biogeographical patterns which are driven by dispersal processes, climate and evolutionary history, such as species-area and distance-decay associations (Horner-Devine et al., 2004; Astorga et al., 2012; Barberán et al., 2015). The field of microbial macroecology has therefore emerged as a promising research path to the unified understanding of ecological processes shaping patterns across different branches in the tree of life.

The contributions of microbiology to macroecology are currently possible largely due to the methodological advances in theoretical and computational tools for investigating microbiomes. Advances in molecular biology and DNA sequencing in the last decade have provided microbial ecologists with new tools allowing the extraction of an unprecedented amount of information from myriads of microbial communities (Snyder et al., 2009). As a result of the growing amount of stored data, new software has been developed for the systematic study of microbial communities on a macroecological scale. Integration among these tools, however, is not a simple task. Microbial macroecology stands to benefit from a formal summary describing the coupling of microbial community characteristics with spatial environmental information.

In this review, we summarize important conceptual challenges as well as computational and methodological opportunities in the study of microbial macroecology, in order to facilitate data integration. We begin by reviewing what has already been described in this field, specifically addressing the conceptual issues of transitioning from micro- to macro- scales when using micro-organisms as model systems. Then, we provide a comprehensive summary of computational tools for describing microbial communities and predicting the distribution of taxa across large spatial scales. Finally, we conclude by proposing a general framework to aid microbiologists in incorporating the peculiarities of micro-organisms into conceptual frameworks of macroecology, in order to promote the unification of microbial ecology and general ecology.

What Have We Done So Far: A Brief Review of Macroecological Studies in Microbiology

Most macroecological studies of microbial communities to date sought primarily to describe patterns in large spatial scales, investigating whether biogeographical patterns exist for the microbiota (Noguez et al., 2005). Most studies were conducted in soil and marine environments and revealed that such patterns do exist. They suggest that environmental predictors for microbiomes could differ from those usually assumed for macroorganisms (i.e., temperature, precipitation and altitude; Fierer and Jackson, 2006); environmental features such as pH, edaphic conditions and land usage are stronger and better predictors for soil microbiomes. However, soil moisture and temperature also appear to be important to predict microbial community composition in some cases (Fierer and Jackson, 2006; Lauber et al., 2009; Drenovsky et al., 2010; Zhou et al., 2016). In marine environments, spatial structure for microbial communities appears to be less prominent (i.e., lower beta-diversity) in comparison to terrestrial and freshwater systems, which is probably due to the more homogeneous abiotic structure of the open ocean (Soininen, 2012) in relation to other environments. Additionally, temperature was a strong predictor for a latitudinal gradient pattern found in planktonic bacteria, with little importance from other variables, such as productivity and salinity (Fuhrman et al., 2008). One study suggested the influence of altitude—a factor that influences that altitudinal patterns of macroorganisms (Lomolino, 2001) —seem to be not relevant for micro-organisms (Fierer et al., 2011). By contrast, Delgado-Baquerizo et al. (2016) stated that altitude gradients are important drivers for microbial diversity considering a wide spatial range (0–4600 m). Finally, it was suggested that micro-organisms in the atmosphere follow a precipitation gradient at continental scales (Barberán et al., 2015). These studies show that some macroecological patterns exist at microbial scales and that they may be similar to those found for macroorganisms in some cases, but not similar in other instances. This raises the question: to which extent are these patterns ubiquitous through all domains of life?

Although much effort has been made to unravel microbial macroecological patterns, so far there is no consensus on which abiotic factors are good predictors of microbial community composition, hampering the implementation of macroecological models to microbial data. Additionally, even though the studies above show strong correlations between variables and microbiome composition, it is still unknown whether the used variables are true drivers of the observed processes, or whether they are actually correlated to unmeasured, confounding factors (Rahbek, 2005). Biotic interactions seem to be equally important in determining community composition; a modeling approach using Artificial Neural Network (Larsen et al., 2012) highlighted the importance of such interactions for creating more accurate models, and a recent study using large microbial community datasets suggested that rarer taxa are better predictors of community structure than environmental factors (Ramirez et al., 2018). Therefore, a modeling framework based on the conceptual idiosyncrasies of microbiomes is required.

Conceptual Challenges for Transitioning Across Spatial and Temporal Scales

An issue arising in all studies addressing microbial macroecology is the proper evaluation of spatial and temporal scales under investigation. The idea that ecological patterns are scale-dependent is pervasive in ecological theory (e.g., Levin, 1992; Crawley and Harral, 2001; Chase and Leibold, 2002; Wu et al., 2002). Two macroecological studies (Willig et al., 2003; Rahbek, 2005) performed at different spatial scales reported distinct patterns for how species richness was associated with latitude and altitude. Hump-shaped patterns dominate species richness and altitude relationships, when the scale of the gradient survey is higher than 1,000 km, but is an uncommon pattern when the scale is below this value. The two studies cited above define two attributes of the sampling design that determine the scale that is being analyzed (**Figure 1**): the unit of sampling and the geographic space covered. The sampling unit is determined by the grain or focus size, i.e., the size of the common analytical unit in the analysis, whereas the geographic space covered, also called the extent, represents the geographical space on which inferences can be made (**Figure 1A**), in other words, the spatial extent covered by all sampling sites (Rahbek, 2005). Macroecological studies investigate processes in large geographical spaces, e.g., continental or global scales (Fierer and Jackson, 2006; Fuhrman et al., 2008; Nelson et al., 2016), which in general define a large extent for macroecological inference. The unit of sampling is represented by the degree of resolution in both response and predictor variables utilized, which can vary widely across studies. Communities' abundance or richness profiles (the response variable) might represent samples in a specific point in space, or samples across different spatial points in the same assumed community (**Figure 1B**). Equally, a single value in a predictor variable (e.g., abiotic conditions, such as temperature, pH, altitude, humidity, etc.) might represent either a 1 km^2 or a 10 km^2 geographic area, depending on how coarse the available environmental information is (Nottingham et al., 2018). The choice and evaluation of the available information is an important step in macroecological studies and may have a deep impact on the results obtained.

Several processes that might be important at local scales may have little effect on, and sometimes even confuse, a pattern at larger spatial scales. For example, Hillebrand (2004) compared studies on the latitudinal species richness gradient, a long well-recognized macroecological pattern, where species richness was

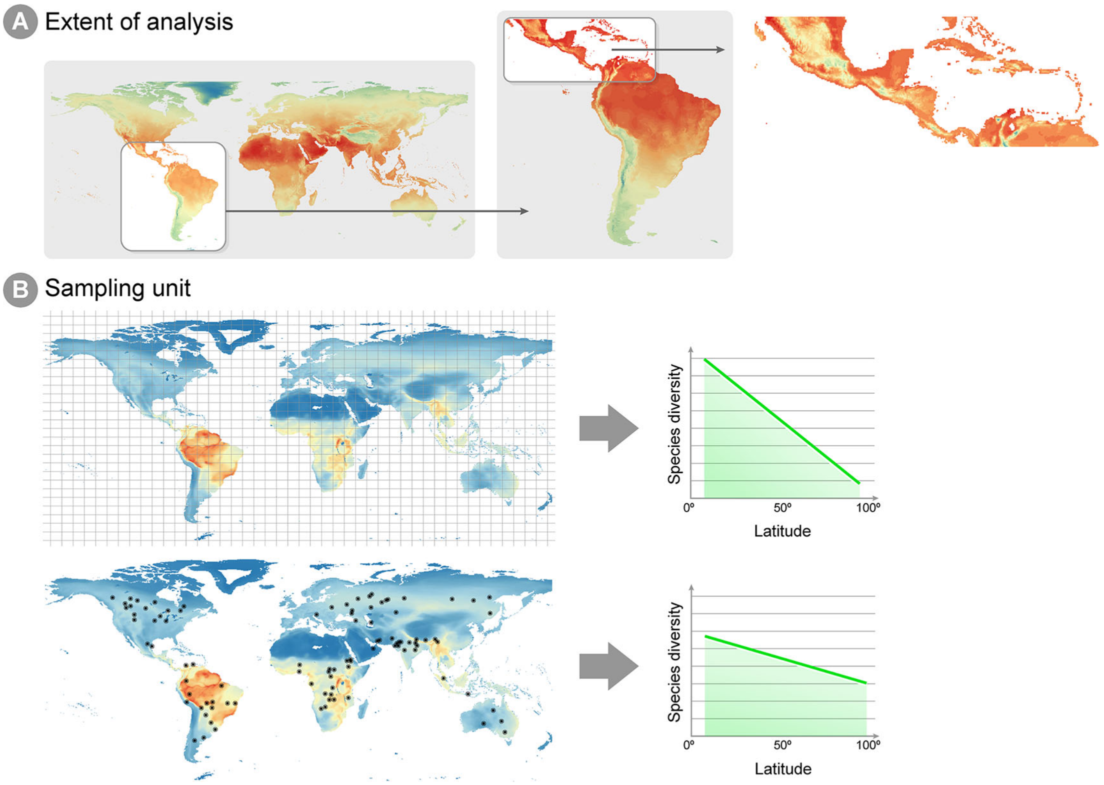


FIGURE 1 | Spatial extent and sampling unit in macroecological analyses. **(A)** Different spatial extents can be analyzed in a macroecological study, which will reflect on the environmental information available for inference and how much extrapolation can be derived from the conclusions of the study. The figure shows annual mean temperature per cell, ranging from low temperatures in blue and high temperatures in red. Notice that the lowest temperatures (blue and green cells) are different for each extent. For instance, when studying Central America, the lowest temperatures can be found in Mexico highlands, whereas an extent focused on the whole Neotropics show lowest temperatures around the Andes mountains. Therefore, caution is necessary when inferences from studies on the Central America are extrapolated to the Neotropics extent. **(B)** Example of two different sampling units in macroecological studies: equally distant squared grids and local sites unevenly distributed through the globe. As highlighted by Hillebrand (2004), squared grids consist of a value averaged across sites within the grid, which decreases the effect of local scale factors (e.g., biotic interactions, dispersal and stochasticity) on the latitude gradient diversity pattern.

known from occurrences equally sized squared areas equally distributed across space (i.e., grids) and studies where species richness was known from sampling points from different studies unevenly distributed across the globe (i.e., local sites). The results demonstrated that the decline of diversity towards higher latitudes was steeper in grid-based studies, suggesting the pattern is easier to detect by using a coarse-grained metric of diversity (as exemplified **Figure 1B**) because local processes (e.g., biotic interactions, dispersal and stochasticity) are averaged out. Additionally, microbial communities seem to be spatially structured mostly at larger study scales (Soininen, 2012), since such scales encompass multiple biogeographical regions separated by dispersal barriers and large variation in climate (Martiny et al., 2006). Therefore, at a smaller spatial scale, community composition may seem stochastic, or greatly vary in short periods of time. The overall conclusion from these studies is that different predictor variables will be biologically relevant at different ecological scales. This suggests that selection a set of predictor variables for model calibration must take into account the ecological scale of the investigated process. Traditionally, in macroecological species distribution models, temperature and precipitation have been successfully used as predictors for macro-organisms, although recent approaches have successfully incorporated biotic interactions into such models (e.g., Araújo and Luoto, 2007; Wisz et al., 2013). A remaining question is whether these same variables are biologically relevant for micro-organisms at large scales. At least for specific and microbiologically diverse ecosystems such as soils, climate—expressed both in terms of climatic factors such as temperature and precipitation, as well as climate-associated attributes such as soil pH, aridity and productivity—is considered a key driver of the structuring and functioning of global microbiomes (Delgado-Baquerizo et al., 2016; Delgado-Baquerizo et al., 2018; Bastida et al., 2019).

There are two main aspects of micro-organisms, which suggest that biologically relevant variables to predict micro-organisms' distribution may indeed be different from those used for macro-organisms. First, micro-organisms exhibit a higher evolutionary rate. Second, due to the organism size, the spatial scale at which micro-organisms perceive the environment is different from that of macro-organisms (Barberán et al., 2014). The first of these aspects indicates that micro-organisms readily adapt to new environments, which means that the distribution range of different microbial taxa is likely to be in equilibrium with environmental variables, which is not always true for macro-organisms (Araújo and Pearson, 2005). Additionally, a high evolutionary rate in micro-organisms indicates that temporal variability in microbiome composition may be high: when environmental changes occur, the microbiome structure is rapidly modified in response, whereas such responses in macro-organisms (expressed in the arrival and disappearance of species, as well as the rise of new adaptations in native species) may take a longer time. This suggests that each microbial sampling is invariably a narrow temporal snapshot of the microbiota, highlighting the importance of time-series sampling to describe for macroecological trends. The very reduced organism size implies that micro-organisms interact with different aspects of the environment, indicating that relevant predictor variables might include, but are certainly not restricted to, large-scale environmental variation. This is still a debatable topic in macroecology of micro-organisms, as some studies argue that micro-organisms respond to continental-scale climatic and environmental variation (e.g., Barberán et al., 2014; Delgado-Baquerizo et al., 2018), whereas others highlight that microscale environmental variation might be more important in predicting distribution patterns (Hendershot et al., 2017). Therefore, when implementing microbiome modeling, one should keep in mind that there is no consensus on which predictor variables should be used. For micro-organisms, the word "environment" might reflect both biotic and abiotic factors surrounding individuals of a species in a defined area, and the relative importance of these two types of factors might be different from what is known for macro-organisms.

The differences between micro- and macro-organisms need to be considered when implementing any of the methods described in this review. For each approach, it is necessary that the macroecological question is clearly stated, and in a way that the scale of sampling and the scale of the studied processes are in agreement with the scale of the proposed questions. In the following sections, we discuss different macroecological approaches for microbiomes, focusing on the description of macroecological patterns and the modeling of microbiomes at macroecological scales. In each case, we highlight how available methods and information can help researchers to answer questions at different spatial and temporal scales.

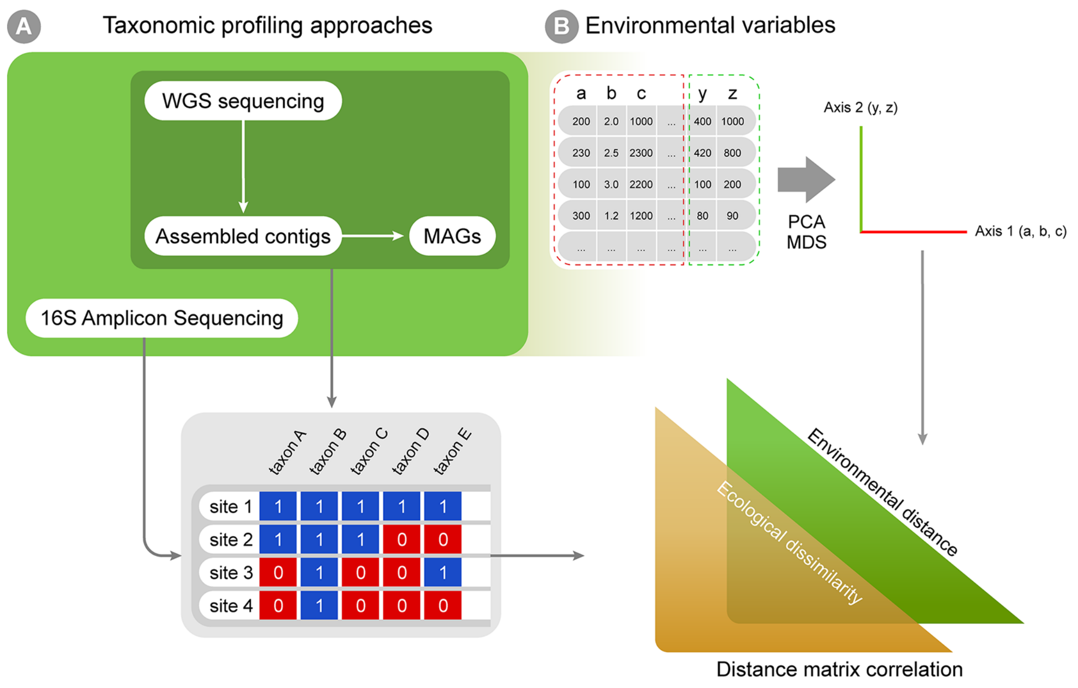
DESCRIBING THE MICROBIOME IN MACROECOLOGICAL SCALES

Taxonomic Profiling and Exploratory Analyses in Microbial Macroecology

The basic input data for macroecological studies is a matrix displaying the presence-absence or abundance data of a biological entity in any taxonomic level across different sampling units (usually a locality defined by a pair of coordinates, but may reflect finer or coarser areas, depending on the specific question, Shade et al., 2018). For microbial communities, such a matrix is usually obtained through the taxonomic annotation of several short DNA sequences (i.e., *reads*) derived from the high-throughput sequencing of an environmental sample (Riesenfeld et al., 2004; Hugenholtz and Tyson, 2008). Reads must first be filtered according to quality and to remove possible contaminants, in order to minimize annotation errors; these tasks can be accomplished using tools such as Prinseq (Schmieder and Edwards, 2011) and Trimmomatic (Bolger et al., 2014). A common and desired practice is to deposit filtered reads in public repositories along with associated metadata, providing public access to the information. This is particularly important for macroecological studies, which make use of secondary data for analysis at large spatial scales. The most prominent repositories for metagenomic data are the NCBI short read archive (SRA; Leinonen et al., 2011b), MG-RAST (Meyer et al., 2008) and the European Nucleotide Archive (ENA; Leinonen et al., 2011a), some of which also provide bioinformatics tools for taxonomic annotation and statistical analysis (e.g., MG-RAST and MGnify; Mitchell et al., 2018). Is worth mentioning that the metadata standard for sequences deposited in International Nucleotide Sequence Database Collection (INSDC) is MIxS (Yilmaz et al., 2011).

Multiple approaches currently exist for obtaining taxonomic profiles from metagenomic sequences, and they mostly fall into four categories depending on the type of data used: 1) amplicon reads, 2) Whole Genome Shotgun (WGS) sequencing reads, 3) assembled contigs and 4) Metagenome-assembled Genomes (MAGs; **Figure 2A**). Each of these has unique advantages and limitations and is suitable to address different scientific questions (**Table 1**). Amplicon analysis consists mostly of PCR amplification of the 16S rRNA gene through the use of degenerate primers designed to cover as much of the diversity of Bacteria and Archaea as possible (Schmidt et al., 1991; McDonald et al., 2012). Next, amplicon sequences are mapped to reference databases, such as RDP (Cole et al., 2014), SILVA (Quast et al., 2013) and Greengenes (DeSantis et al., 2006), which contain pre-computed high-quality alignments of 16S rRNA genes, allowing for fast taxonomic assignments for millions of sequences. This approach tends to be accurate at low taxonomical levels (e.g., genera) and is cost effective, considering the coverage of sequencing per sample, making it possible to sample many more replicates per study. On the other hand calculating taxa abundances across samples can be a limitation due to the presence of multiple copies of the 16S rRNA gene in a single genome. Additionally, the so-called universal primers used for amplicon analysis usually do not amplify genes derived from major fractions of the diversity of Bacteria and Archaea, such as the candidate phyla radiation (Hug et al., 2016a).

One common alternative to amplicon sequencing is Whole Genome Shotgun (WGS), i.e., the sequencing of DNA fragments covering the whole diversity of genes in an environmental sample. Similar to amplicon based studies, WGS reads are annotated by comparing them to previously characterized sequences deposited in reference databases, encompassing genes from multiple taxa. This comparison can be based on homology or the search for similar k-mer profiles (i.e., the set of all possible sub-strings of different lengths for a DNA sequence). Due to redundancy in the genetic code, proteins are more conserved than nucleotide sequences; using homology to detect similar protein sequences is more sensitive and suitable for detecting distant evolutionary relationships, allowing more sequences to be classified. Because the degree of identity between the sequences of naturally occurring microbes and those available in reference databases is often very low, annotations of WGS reads often require using permissive cutoffs (i.e., reads are assigned to a taxon even if the identity is



a	b	c	...	y	z
200	2.0	1000	...	400	1000
230	2.5	2300	...	420	800
100	3.0	2200	...	100	200
300	1.2	1200	...	80	90
...	...	...	...	...	...

	taxon A	taxon B	taxon C	taxon D	taxon E
site 1	1	1	1	1	1
site 2	1	1	1	0	0
site 3	0	1	0	0	1
site 4	0	1	0	0	0

FIGURE 2 | A workflow summary for taxonomic annotation and exploratory analyses. Taxonomic annotation methods are used to generate, for instance, presence-absence matrices **(A)**, which can be combined with environmental variables into correlation analyses **(B)**. The biological variation in environmental variables can be simplified through ordination analyses (such as PCA and MDS). Finally, distance matrices can be created for both ecological and environmental variation, and distance matrix correlation can be used to infer if environmental distances correlate with ecological differences among sampling sites.

TABLE 1 | Approaches for obtaining taxonomic profiles from metagenomic samples.

Input type	Software	Speed	Reference Databases	Confidence	Advantages
Amplicon	Qiime, MOTHUR	Fast	SILVA, RDP and Greengenes	Low	Extensive databases of sequences and samples for comparison
WGS Homology	Diamond, BLAST, BLAT, MEGAN	Slow	nr, Uniprot, pfam	Medium	Based on the whole genetic diversity
WGS K-mer	Kraken, FOCUS	Fast	RefSeq Genomes	Medium	Based on the whole genetic diversity
Assembled Contigs	Assembly: SPAdes, IDBA_ud, Ray-Meta Contig Classification: CAT, MEGAN, Kaiju	Slow	nr, Uniprot, pfam	High	Discovery of new taxa, more reads assigned
MAG	Assembly: SPAdes, IDBA_ud, Ray-Meta Binning: Metabat, GroopM, ABAWACA, CheckM Classification: CAT/BAT	Slow	N/A	High	Yields draft or complete genomes, discovery of new taxa, more reads assigned

low, e.g., only 30%), provided that it falls within other assumed cutoffs of alignment, length and e-value. Several reference databases are currently available, as well as tools to detect protein-protein and protein-nucleotide homology (**Table 1**). As an alternative to homology searches, k-mer composition profiles are significantly faster and make it possible to rapidly analyze a large number of samples (**Table 1**).

Using WGS sequencing further allows for the assembly of raw reads into larger contigs, and, in some cases, later binning into metagenome-assembled genomes (MAGs; **Figure 2**). This approach may improve taxonomic classification by assessing longer genomic fragments that derive from such sequence assembly. The Critical Assessment of Metagenome Interpretation (CAMI) challenge reviewed several metagenomics tools (Sczyrba et al., 2017). This study distinguished between taxonomic binners (which allow taxonomic abundances to be inferred by clustering individual sequences, then assessing longer genomic fragments Lin and Liao, 2016; Wu et al., 2016), from taxonomic profilers (which focus on predicting a taxonomic abundance profile without necessarily classifying every sequence, often assessing only raw reads Ounit et al., 2015; Koslicki and Falush, 2016). They show that classifiers in general were more accurate than profilers in estimating the relative abundances of taxa. This increased performance is due to the fact that longer sequences contain more phylogenetic information than short reads, leading to less noise in the taxonomic profile. Moreover, because sequence assembly reduces the total volume of sequence data to be classified, more sensitive homology searches that are computationally more demanding may be applied than the rapid searches that are used for classification of short, raw reads. Two recently developed tools that explicitly exploit the added information in assembled contigs are MEGAN-LR (Huson et al., 2018) and the Contig Annotation Tool [CAT, (von Meijenfeldt et al., 2019); https://github.com/dutilh/cat] that exploit all sequences in the full GenBank reference database for taxonomic classification. A limitation of metagenomic assembly is that it is susceptible to possible errors arising during the assembly, which is aggravated when population diversity of the sampled microbial community is high (Sczyrba et al., 2017). Moreover, high levels of sequence heterogeneity between related strains may lead to abundant genomes in the sample being misassembled as chimeras, and potentially misclassified. The subsampling of shotgun metagenomic reads before assembly has been applied to resolve this problem (Hug et al., 2016b).

Once contigs have been assembled into longer fragments of the genomes present in the community, metagenome-assembled genomes (MAGs) may be reconstructed by binning contigs from the same genome together. Several software tools are available to perform MAG reconstruction (**Table 1**). At this stage, phylogenetic and phylogenomic methods can be used to determine the taxonomic affiliation of these MAGs with even more confidence than that of individual contigs. Additionally, MAGs and assembled contigs can be used to build custom sample-specific reference databases for read mapping (e.g., Speth et al., 2016). The main advantage of using such databases is that often many more reads can be assigned, because the contig sequences represent the strains that are reconstructed from the same sample, minimizing the occurrence of false positives. Therefore, the obtained taxonomic profile contains less noise and more comprehensively represents the data.

The taxonomic profiles obtained from the methods above can be assembled into presence-absence or abundance matrices and further explored using classic multivariate exploratory analyses, such as multivariate ordination/canonical methods (Hanson et al., 2012; Xue et al., 2018). Under the macroecological rationale, exploratory analyses are used to describe the biological variation across a global or continental gradient in potential explanatory variables (e.g., describing diversity or abundance variation across the latitudinal temperature gradient, continental atmospheric variation, etc.; Shade et al., 2018). Correlation among explanatory variables is a common issue in biological statistics, and multivariate ordination is then used to reduce dimensionality and yield new mathematically uncorrelated axis from the original correlated explanatory variables (Legendre and Legendre, 2012; **Figure 2B**). A few approaches widely used for this purpose are: 1) Principal components analysis (PCA), which is based on covariance or correlation matrices and is suitable for sets of linearly correlated measures; 2) principal coordinates analysis (PCoA), which differs from the PCA by extracting eigenvalues from similarity or distance matrices, therefore being appropriate for non-linear relationships; 3) multidimensional scaling (MDS) that, unlike PCA and PCoA, is not based on eigenvalues decomposition and, like PCoA, is limited to Euclidean distances matrices and 4) correspondence analysis (CA), based on contingency table of categorical variables (Bray and Curtis, 1957; Clarke, 1993). The new mathematical axes provide a mathematical space where measurements from the actual environmental samples can be placed and compared. The associations between variables (e.g., diversity and temperature) can also be tested by classic statistical analyses like regression and correlation, which can be based on both original explanatory variables and new mathematical axes created by ordination analyses. Additionally, ecological similarity between localities can be explored using distance measures (e.g. Euclidean, Mahalanobis, Jaccard, and Bray-Curtis) and compared against a distance matrix for a potential explanatory variable in the same localities and statistical significance can then be assessed by using a test such as the Mantel test (**Figure 2B**). Such approaches are commonly used in macroecological studies to statistically assess the correlation between two distance matrices based on variables of interest (e.g., Duarte et al., 2009; Bell, 2010).

Describing Community Structure With Co-Occurrence Networks

Co-occurrence networks (CNs) has been used to describe associations within microbial community (**Figure 3**). Usually, in these networks, the nodes represent taxa and the edges represent statistically significant positive or negative correlations in the abundance of taxa across several samples in a given environment or host (Faust and Raes, 2012). A few authors may also include abiotic factors as nodes (e.g. Li et al., 2017). Using CNs can reveal insights about possible ecological interactions and distribution patterns of microbial taxa (Faust and Raes, 2012; Cardona et al., 2016). Two important types of information can be retrieved from CNs: 1) changes in community structure across environmental gradients, that is, variation not only in the species abundance, but especially in the degree of correlation between taxa across environmental gradients; and 2) potential biotic interactions that can be useful

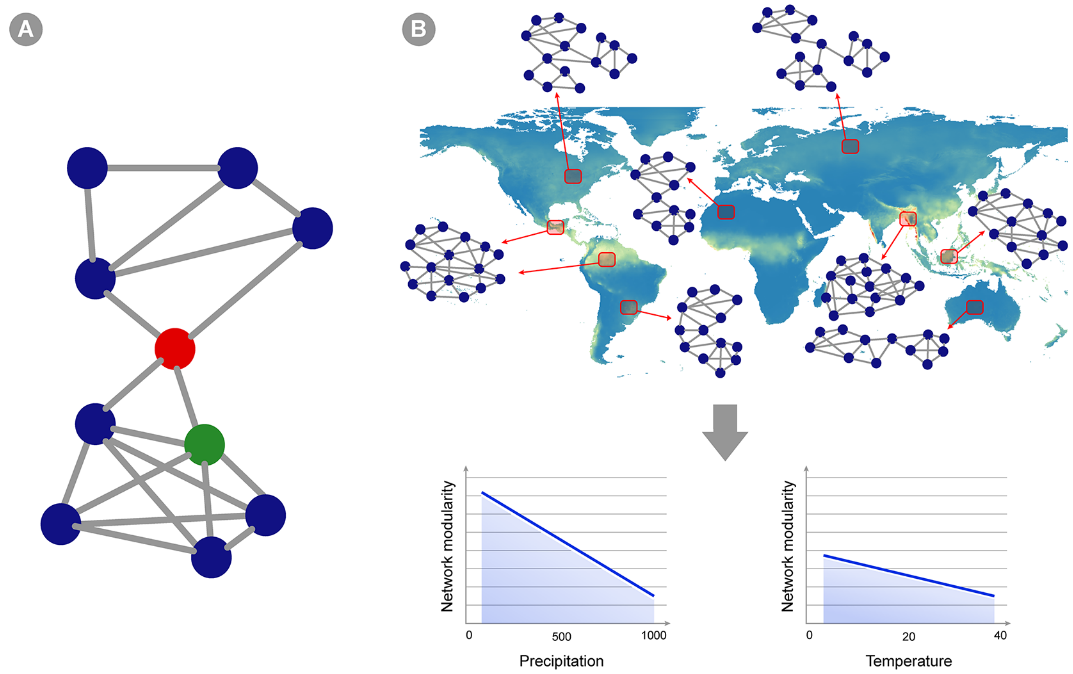


FIGURE 3 | Co-occurrence networks applied to microbial macroecology. **(A)** A hypothetical example of a co-occurrence network. Circles represent different taxa and edges connecting two circles indicate statistically significant co-occurrence between those two taxa, i.e., they co-occur more than expected by chance in the set of samples analyzed. Network structure can indicate ecosystem properties, and these can be translated into statistics summarizing network topology (see Box 1). For instance, this hypothetical network shows two subunits (or modules) separated by the taxon indicated as a red circle. This taxon is also a node with high betweenness centrality (i.e., indirect connections between any two nodes in the network has a high probability of going through this node), whereas the green circle represents a node with high degree (i.e., showing a connection to many other taxa). **(B)** A hypothetical example of a macroecological study using co-occurrence networks. Red squares represent an area where several samples were gathered and analyzed, yielding a single abundance matrix and a corresponding co-occurrence network (two sites pointing to the same network represent areas in which networks are highly similar). The topology of the network changes in different ecosystems across the globe, and the overall hypothetical pattern is represented in the graphics below: network modularity (i.e., defined as the number of subunits within the network, as well as the relative proportion between connections within and between modules) decreases as precipitation and temperature increases (but the change is less intense for temperature).

for macroecological modeling (*Predicting Microbial Distribution and Community |Composition*). Since CNs are based on abundance correlation, it is desirable that they are built over a large number of sampling units, and therefore hold great potential for application in macroecological studies (Berry and Widder, 2014). Distinct approaches have been used to construct CNs and derive information from their structure, such as distance or similarity matrix metrics among the samples used to construct the networks (Fan et al., 2018; Jackson et al., 2018; Marasco et al., 2018; **Box 1**). Overall, the same matrix generated by the software tools listed in the previous section can be used as input for CN calculation. Samples can be grouped according to the macroecological variable of interest (e.g., temperature variation across latitudes, atmospheric variation across a continent, variation in land cover across the globe) and the structure of CNs from each of these groupings can be compared across global or continental scales (**Figure 3**). Note that comparison of microbial community structure has often been performed across different ecosystems (e.g., comparing the structure of networks between fresh and saline water environment), but the macroecological approach supports the rationale of a comparison within the same environment (e.g., soil samples) across an environmental gradient (e.g., temperature, pH, etc.; Barberán et al., 2012). Several measures exist to describe network structure, such as symmetry, degree distribution, checkerboard index (Horner-Devine et al., 2007; Araújo et al.,

BOX 1 | Building and Interpreting Co-Occurrence Networks.

Several tools are available to build and interpret co-occurrence networks. The software CoNet (Faust and Raes, 2016), developed in Cytoscape (Shannon et al., 2003), allows the usage of several measures for dependency, similarity and dissimilarity, to build and visualize co-occurrence networks. In order to build these CNs, the microbial composition data is provided in relative abundances. Some annotation tools provide microbial composition in read counts, in this case one can use SparCC (Friedman and Alm, 2012), which calculates abundance correlations among taxa without the issues associated with compositional data (Mendes et al., 2018), for further CNs analysis. Alternatives to SparCC are REBACCA (Ban et al., 2015) and CCLasso (Fang et al., 2015). Kurtz et al. (2015) presented another tool: SPIEC-EASY, a pipeline that transforms relative abundance data and estimates interaction graphs. Finally, a few approaches are based on information theory, for instance: using mutual information combined with other metrics, implemented in CoNet (Lima-Mendez et al., 2015). Choosing a correlation method for network construction is critical once networks generated by different methods can provide contrasting results (Weiss et al., 2016). Methods should be chosen taking into consideration if microbial community data are presented as relative abundance or in absolute read counts.

Keystones in CN

There is no consensus on the operational definition of keystone for microbial ecology (reviewed in Banerjee et al., 2018). However, a usually proposed definition is that keystones are highly connected microbial taxa presenting a unique and crucial role for community structure and functioning, so their loss or removal should have large impacts on microbial community (Banerjee et al., 2018). In this sense, network theory provides us with quantitative ways to characterize how connected a given microbial taxa is. One criterion, based in network theory, to determine a putative keystone taxon is high betweenness centrality (BC; e.g., Lupatini et al., 2014; Banerjee et al., 2016; Jiao et al., 2016; Li et al., 2017; Mendes et al., 2018), albeit an investigation based on dynamical modeling found lower BC to be correlated with higher probability of a taxon being keystone (Berry and Widder, 2014). The BC of a node A is the number of shortest paths connecting two nodes which pass through the node A. Nodes with high BC connect portions of the network that would otherwise be sparsely or not connected at all. Therefore, removing high BC nodes leads to a sparser network, disconnecting modules in several cases. The number of connections a node presents, which is called the node's degree, is also a frequent metric used as a keystone index (Comte et al., 2016; Hartman et al., 2018). This is based on the idea that, taxa (nodes) that are connected with multiple others are important to network structure, and their potential removal would have a high impact to the community. It is interesting to highlight that, whereas one node can have both high degree and high BC (in which case this taxa would be considered keystone by both definitions), it is also possible to find nodes in which BC is high and degree is low, or vice-versa, leading to a disagreement between these two keystones definitions. Therefore, it is important to have in mind the biological process of interest because this will determine the more important features in a given community and what keystone definition one should use.

A different approach, based on metabolic networks (Guimera and Amaral, 2005), assumes that the network is formed by modules (i.e., semi-independent groups of cohesive, interacting taxa). In this approach, one can calculate the z-score, which is a measure of the number of interactions a taxon has within its module; and the c-score, which describes how evenly distributed are the interactions of a given taxon across multiple modules. These two values allow us to classify the taxa in network hubs (z-score > 2.5; c-score > 0.6), module hubs (z-score > 2.5; c-score < 0.6), connectors (z-score < 2.5; c-score > 0.6) and peripherals (z-score < 2.5; c-score < 0.6) (Poudel et al., 2016; Fan et al., 2018). Putative keystones taxa would then be the nodes identified as network hubs, module hubs and connectors. One advantage is that this definition takes into account multiple features that might make a node important to a network (e.g., participating in a network within a hub or as connectors between hubs), whereas, when one looks only at BC or high degree, a single type of keystone feature is taken into account.

Indirect Effects From CNs

In networks, species that do not directly interact can influence each other through cascading effects that spread through the network (indirect effects). Guimarães et al. (2017) developed an analytical framework to quantify the total amount and the importance of the indirect effects in a given network. Their results show that network structure is what drives how the indirect effects spread through the network (Guimarães et al., 2017). Networks of micro-organisms, which are species-rich networks formed by a small core of highly connected species and many species poorly connected (Banerjee et al., 2018), are predicted to show a higher amount of indirect effects than poor, highly modular networks. Therefore, quantifying indirect effects might be an important aspect in the study of which micro-organisms are keystones to a given community relevant to maintain relevant ecosystems functions and contribution to resilience and stability in face of global environmental changes (Berry and Widder, 2014).

In addition to measuring indirect effects, it is possible to explore the consequences of such effects. Resilience and stability are important aspects of network structure that can be measured by using approaches derived from the study of dynamical systems. Coyte et al. (2015) proposed an extremely general and suitable framework that can be used to analyze species-rich microbial networks. Their approach uses the eigenvalues of the matrix that describes the effects of ecological interactions at the equilibrium (Jacobian matrix) associated to a given network, to analyze the stability and resilience of microbiome networks. Their approach can be used in networks that possess any combination of different types of interactions (cooperation, competition, exploitation, amensalism and commensalism). One important result of their analyses is that cooperation tends to destabilize microbial networks. The destabilization effect happens because of the presence of positive feedbacks between the species when they cooperate, which leads to cascading effects. For example, a decrease in population size of one species might lead to all the species they positively interact with to decrease as well. On the other hand, competition gives a stabilizing effect in the network; compensating the destabilizing effect that increasing richness can have in an ecological community (May, 1972).

2011; Layeghifard et al., 2017), but the best usage of such metrics is an ongoing debate (Layeghifard et al., 2017) and is highly dependent on the ecological question being asked.

Co-occurrence networks may also be used to identify keystone taxa (**Box 1**). The keystone concept was first coined by Paine (1966), who demonstrated that the removal of the sea-star *Pisaster ochraceus* caused a dramatic change in community structure on a rocky shore, concluding that the species functioned as an important element for maintaining community integrity, most likely due to its non-redundant role (Paine, 1969). This definition can be applied in the microbial ecosystem and be empirically investigated by using network approaches. Keystone taxa can be compared across macroecological scales to investigate whether and how the importance of specific groups as key taxa in communities across an environment varies on global scales. Since keystone taxa usually perform important and non-redundant functions, their identification may be important to understanding ecosystem functioning.Thus, an approach coupling keystone identification with measurements of functional diversity across macroecological scales holds potential to bring numerous insights (see below). Finally, another insight derived from CNs is how the network structure may favor or constrain cascading effects (**Box 1**), which may favor or imperil the resilience of the communities against perturbations (another ongoing debate within ecosystem ecology; Oliver et al., 2015). Cascading effects often propagate across networks, connecting the dynamics of taxa that do not directly interact with each other. In fact, networks of taxa are subject to influences from taxa they directly interact with, as well as to indirect effects that pervade the network, i.e. from taxa with which they do not interact directly. Under certain conditions, indirect effects can be more important to the network dynamics than the direct effects (Ohgushi, 2005). Indirect effects can be measured across macroecological scales to assess, in a spatially explicit manner, in which ecosystems indirect effects seem to play a more important role to maintain microbial community stability (Guimarães et al., 2017).

Revealing Macroecological Patterns From Microbiome Functional Diversity

Functional ecology, defined as the study of the roles that organisms play in their ecosystems, also holds great potential for microbial macroecology. Studies investigating levels of functional diversity across macroecological scales are already common for macro-organisms (Fu et al., 2017; Jarzyna and Jetz, 2018), both in theoretical investigations of processes determining functional diversity (Safi et al., 2011) and in more practical inquiries such as the conservation of ecosystem functions (Devictor et al., 2010). Yet similar studies have not been performed for micro-organisms. For instance, previous studies have explored like global patterns of mammalian functional diversity (Safi et al., 2011) as well as global scale marine macroecological patterns (Amend et al., 2013) have no equivalent investigation concerning microbial functional diversity. Macroecological studies might yield insights on the patterns observed for the functional diversity of micro-organisms across different environments in the globe, and address their relation to ecosystem functioning and service provision (Mace et al., 2012).

Functional diversity is one of the three main biodiversity dimensions investigated in macroecology, alongside taxonomic and phylogenetic diversity (Webb et al., 2002; Devictor et al., 2010). Functional diversity is usually defined as the amount, variation and distribution of traits in a community (Dıaz and Cabido, 2001), originally measured by the calculation of the total branch length of the functional dendrogram constructed from information about taxa' functional traits (Petchey and Gaston, 2002). From this initial method, several new conceptual and mathematical approaches have been developed and implemented (a few revised in Petchey et al., 2004), but none of them dismiss the need to 1) choose the functional traits through which organisms will be distinguished, 2) define how the diversity of the trait information will be summarized into a measure of functional diversity, and 3) validate the measurements through quantitative analyses and experimental tests (Petchey and Gaston, 2006). In micro-organisms, functional traits are usually viewed as the genetic and biochemical characteristics of organisms affecting ecosystem functioning, such as the production of metabolic inhibitors or enhancers, or enzymes playing a role in ecosystem metabolic pathways (Dıaz and Cabido, 2001). In this sense, the function of micro-organisms in an ecosystem is defined by their genetic composition, which ultimately dictates the molecules they metabolize (Faure and Joly, 2016). Similar to taxonomic annotation, functional traits can be derived by direct functional annotation of metagenomic short-reads from an environmental sample (with no taxonomic annotation). Alternatively, prior metataxonomic approaches (e.g., 16S rRNA) can be used to taxonomically assign individuals in a sample, and then functional annotation can be derived from their phylogenetic position. Software tools to perform both approaches are summarized in **Table 2**, with their respective references. All of these metagenomic and metataxonomic functional annotation approaches are based on genomic databases and the accuracy of annotation depends on the quality of software databases. Furthermore, many genes are still unassigned, and their functions are unknown, making it challenging to infer ecological functions from genetic content alone (Faure and Joly, 2016).

The degree of functional diversity has been used to investigate two main macroecological patterns in microbial communities: 1) relationships between community taxonomic and functional composition among microbial communities (Louca et al., 2016; Vieira-Silva et al., 2016; Galand et al., 2018) and; 2) how microbial functions vary in time and space (Dinsdale et al., 2008; Ren et al., 2017; Galand et al., 2018). Usually the most accessed functional measures are diversity (including functional richness, evenness and divergence), composition, redundancy and rarity. Several algorithms and computational tools have been published in order to assess and quantify these functional features (**Table 3**, also reviewed in Mouchet et al., 2010; Schleuter et al., 2010; Song et al., 2014; Bond-Lamberty et al.,

TABLE 2 | Tools used to annotate functional potential profiles from metagenomic reads or to infer them from 16S taxonomic annotation.

Tool	Approach	Synopsis	Features	Reference
BLASTx	Read annotation	Uses alignment approach to annotate nucleotide reads into potential proteins	+ great sensitivity - it can be very slow for high-throughput data	Altschul et al. (1990)
MetaGeneAnnotator	Read annotation	Identify putative proteins by estimating di-codon frequencies through the GC content of a nucleotide read	- not precisely estimate de Domain of a given sequence	Noguchi et al. (2006)
DIAMOND	Read annotation	Uses double indexing alignment to annotate nucleotide reads into potential proteins	+ 2000 to 20000 times faster than BLASTx	Buchfink et al. (2015)
SUPER-FOCUS	Read annotation	Functional profiling of metagenomes	+ output consists in a three hierarchical level functional profile, useful to choose your level of functional resolution	Silva et al. (2016b)
MGS-Fast	Read annotation	Preprocess and analyses WGS reads into functional profiles by using stringent DNA-DNA matching to the IGC database.	+ includes preprocessing steps (read trimming and removal of low-quality sequences) and taxonomic profiling	Brown et al. (2019)
MetaCLADE	Read annotation	Uses a multi-source domain annotation strategy to profile reads into protein domains.	+ designed to also annotate metatranscriptomic reads	Ugarte et al. (2018)
PICRUSt	16S inference	Uses evolutionary modelling to predict community putative functional profiles from 16S marker gene using a genome reference database	+ online interface to users unfamiliar with programming	Langille et al. (2013)
PAPRICA	16S inference	Places reads into a 16S phylogenetic tree of consensus genomes to predict the functional profile	+ very accurate to infer functional profile of well-known organisms that have plenty of genomes in the database	Bowman and Ducklow (2015)
FAPROTAX	16S inference	Extrapolates community taxonomy into putative functional profiles	- database used from cultivated organisms only	Louca et al. (2016)
QIIME	Functional pipeline	Provides a wide range of microbial assembly analysis and visualizations from raw nucleotide sequences	+ network and phylogenetic analysis and core assessment	Caporaso et al. (2010)
MOCAT2	Functional pipeline	Assemble and quality-filter reads to comprehensively predict them functionally and quantify them	+ also annotate metagenomes taxonomically	Kultima et al. (2016)

2016; Ricotta et al., 2016). Addressing the above-cited questions, one of the emerging patterns in micro-organisms is a decoupling between functional and taxonomic composition (Louca et al., 2016). This trait suggests that microbial communities may present a high degree of functional redundancy, meaning that shifts in taxonomic community composition do not lead to shifts in functional community composition. It has been hypothesized that the mechanisms underlying microbial assemblage are distinct from mechanisms governing functional composition, and that environmental factors are potential predictors of functional composition (Louca et al., 2018). We further suggest that approaches for characterizing functional diversity should also be coupled with estimates of function turn-over and nestedness; metrics that in macroecology are commonly used to measure shifts in species composition mostly along abiotic gradients, the so called *beta*-diversity (Legendre et al., 2005; Anderson et al., 2006; Jost, 2007). This information would allow us to answer questions such as whether a specific subset of

TABLE 3 | Tools to calculate functional diversity features.

Tool	Approach	Synopsis	Features	Reference
PHYLOCOM	Software	Calculates trait distribution to compare with random community consortia as well as uses evolutionary models to simulate trait and phylogenetic evolution	+ uses null models to test hypothesis of trait similarity + integrates trait information with evolutionary analysis + able to deal with polytomies	Webb et al. (2008)
FDiversity	Software	Focuses on calculation of functional diversity indexes and statistically analyze them	+ user friendly interface + accepts different input data formats	Casanoves et al. (2011)
FD	R-language package	Uses functional dispersion index and measures diversity based on distances of traits in a multidimensional space	+ allows missing values on calculation + allows weighting traits per abundance	Laliberté and Legendre (2010)
SYNCSA	R-language package	Uses matrix correlation to estimate trait patterns, phylogenetic signal and environmental variations for metacommunities	+ allows environmental characteristics to be considered	Debastiani and Pillar (2012)
cati	R-language package	Estimates community assembly patterns by species interactions and environmental filtering	+ allows differentiation among individuals + can integrate phylogenetic information into analysis	Taudiere and Violle (2016)
funrar	R-language package	Estimates functional rarity based on abundance and/or spatial frequency of species	+ estimates functional uniqueness, distinctiveness and taxon scarcity and restrictedness	Grenié et al. (2017)

functions is filtered and maintained in a specific environment; or how functions are changing across abiotic gradients.

PREDICTING MICROBIAL DISTRIBUTION AND COMMUNITY COMPOSITION

Macroecologists describe spatial patterns of biodiversity aiming to ultimately create accurate models that can predict biodiversity under different scenarios. The patterns described are analyzed, and the underlying biotic and abiotic drivers of species distribution and abundance are tested in a statistical framework. Understanding the mechanisms behind these patterns allows macroecologists to predict biodiversity in geographic areas not yet studied, contributing to decrease biodiversity shortfalls (Hortal et al., 2015) as well as how biodiversity would respond to changes in the environment (Kerr et al., 2007). The BAM (as an abbreviation for '*biotic, abiotic and movements*') diagram is a conceptual framework used in macroecological modeling to summarize the determinants of species distribution on global scales (**Figure 4**; Soberón and Nakamura, 2009).

In the BAM framework, the presence of a focal species in a specific site is determined by: (1) the presence, absence and/or abundance of other species in the same environment (i.e., biotic factors, the B in BAM); (2) the availability of the environmental attributes that are suitable for the focal species (i.e., abiotic factors, the A in BAM) and; (3) the focal species capacity to migrate into biotically and abiotically suitable areas (i.e., movement capacity, the M in BAM; **Figure 4**). This idea is described in a more formal manner in the Hutchinsonian concept of ecological niche, i.e., the n-dimensional hypervolume in which a species can exist (Colwell and Rangel, 2009; Holt, 2009; **Figure 4**). This conceptual framework is important for models that attempt to predict the occurrence of taxa, since it highlights which factors are expected to affect taxa presence in different locations. For macroorganisms, models are usually calibrated with the usage of abiotic factors at large spatial scales, specifically temperature and precipitation, which were shown to be good predictors of terrestrial species distribution range (e.g., Soberón, 2010). Such models usually show acceptable accuracy, but several studies highlight the importance of accounting for migration capacity and species interactions in distribution modeling (Araújo and Luoto, 2007; Wisz et al., 2013).

When it comes to micro-organisms, it is necessary to clearly understand which factors affect the distribution of microbial species. The BAM diagram offers an adequate conceptual framework to start addressing this question. Several authors

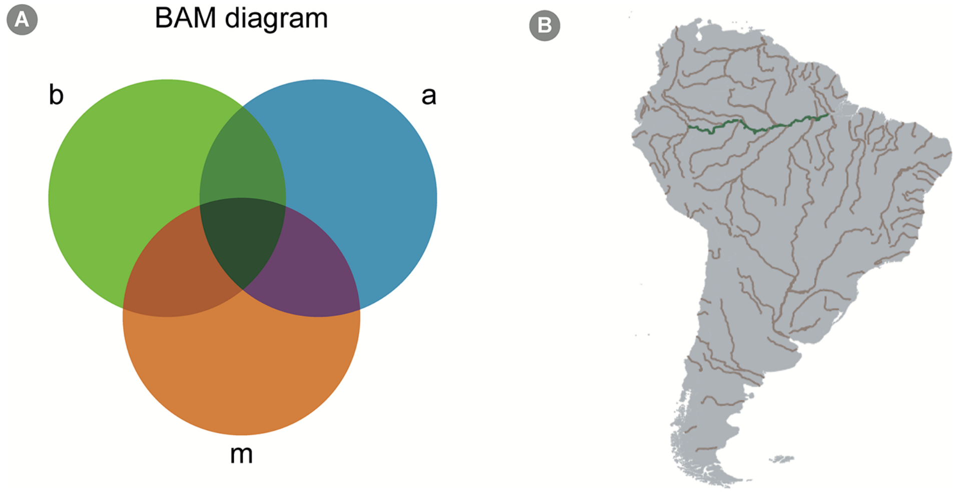


FIGURE 4 | The BAM Diagram. **(A)** A scheme of a hypothetical BAM diagram (abbreviation for "*biotic, abiotic, and movements*"), highlighting the intersection between the different aspects determining the presence-absence of species. The *b* circle, colored in green, represents biological aspects allowing the presence of the species; the *a* circle, colored in blue, represents the abiotic aspects; finally, the *m* circle, colored in orange, represents the movement aspect, which consists in the dispersal capacity of the species. The intersection represents areas where more than one of those aspects allows the existence of the species. For instance, the green intersection represents an area where both biotic and abiotic conditions allow the species to exist, but the species is unlikely to disperse to that area. Similarly, the purple intersection represents an area where abiotic conditions allow the species to exist and is within the species' dispersal capacity; however, biotic conditions (for example, presence or absence of important species with which it interacts) do not allow their existence. All species occur only in areas represented by the dark green intersection, i.e. the intersection of all three factors. Mathematical models, however, can calibrate species niche based, solely on abiotic factors (which is the case of most SDM approaches), and, in these cases, the BAM diagram is a good conceptual framework to interpret the results. **(B)** A geographical projection of the BAM diagram for a hypothetical microorganism in South America. The grey areas across the continent represent sites to where the species can potentially disperse to (based on the idea that micro-organisms have high dispersal capacity, see *Predicting Microbial Distribution and Community Composition* in text). Assuming our hypothetical species prefer freshwater conditions, rivers in South America are colored in brown, to represent the intersection between factors *a* and *m* in the diagram. Finally, the green color of the Amazon river indicates an area where all factors allow the existence of the species (i.e., the species can disperse to the area, it is a freshwater environment, and it shows biotic conditions favorable to its establishment, e.g. the presence of specific species with which it cooperates).

have suggested that the dispersal capacity of micro-organisms is much higher than that of macroorganisms (Finlay and Clarke, 1999; Martiny et al., 2006; Barberán et al., 2014). In this aspect, the movement feature of the BAM diagram would have little effect on the distribution of species, since several studies indicate that micro-organisms are highly dispersive (Bovallius et al., 1980; Fenchel and Finlay, 2004; Martiny et al., 2006; Barberán et al., 2014; but see, e.g., Peay et al., 2010), and that spatial structuring of microbial communities are only perceivable on large spatial scales. This leaves us with the biotic and abiotic factors as major drivers of micro-organisms' distribution. As previously discussed in *Conceptual Challenges for Transitioning Across Spatial and Temporal Scales*, a few studies have highlighted the importance of different abiotic factors in structuring microbial community, which are not always related to the environmental predictors used in distribution modeling of macroorganisms. Such variables include, besides temperature and precipitation, edaphic conditions, soil pH and concentrations of different chemical molecules (Lauber et al., 2009; Drenovsky et al., 2010; Zhou et al., 2016). Additionally, the biotic interactions among species have been advocated as important determinants of species occurrence (Larsen et al., 2012; Ramirez et al., 2015; Ramirez et al., 2018). Therefore, in the following sections we describe how to access available spatial-explicit environmental data for micro-organisms modeling, as well as modeling approaches that can account for both biotic and abiotic factors.

Using Abiotic Variables to Model Microbial Communities

Each sample taken from the environment is under the influence of a huge number of variables in many spatial and temporal scales. In order to model the composition of microbiomes, and therefore the distribution of micro-organisms across the globe, it is important to have available environmental data on the relevant spatial and temporal scales. The variables used to model micro-organisms will depend on the specific environment under study. Micro-organisms living in the soil are affected by different environmental factors than those living in a freshwater lake or in the ocean. This is different than what is seen for macroorganisms, where global temperature and precipitation play major roles defining biogeographic realms (McGill, 2010). While acknowledging that global variation in temperature and precipitation might define biogeographic areas for micro-organisms (Martiny et al., 2006), we argue that this definition will differ when comparing between micro-organisms living in different environment types (e.g., soil vs freshwater micro-organisms).

Physical properties are usually important in several environments, such as temperature, precipitation, moisture and solar radiation. These variables can be measured or modeled *via* remote sensing platforms and remote sensing-based modeling tools. Due to the advent of environmental monitoring satellites and the creation of on-line data processing and distribution platforms, there is a wealth of environmental data with global coverage available to the general public, ranging from raw satellite images to validated measurements of parameters, such as land surface temperature, precipitation rates, the concentration of gases such as CO_2 in the troposphere and photosynthetic activity (**Table 4**). These databases contain climatic spatially explicit information such as land surface temperature, net primary productivity, vegetation and leaf area indexes, evapotranspiration, detailed landcover map and precipitation rate. Additionally, since other aspects of soil and atmosphere might also be necessary to fully characterize the abiotic environment of micro-organisms. Information pertaining to soil physical (e.g., clay content) and chemical (e.g., pH) conditions, as well as soil classification across the globe can be retrieved from these databases. Similarly, when investigating the atmosphere microbiome, the atmospheric chemical composition may play a large role on community composition by changing the chemical properties such as pH and playing an important role on ecological processes, such as nitrification (Keller et al., 2006; Hutchins et al., 2009; Hatzenpichler, 2012). An example of atmospheric chemical composition data available, such as the products based on the Atmospheric Infrared Sounder (AIRS), is a hyperspectral instrument on board of Aqua satellite (**Table 4**). By decomposing the infrared radiation in 2,378 bands, AIRS can provide daily measurements of trace components abundances in the atmosphere, including ozone, carbon monoxide, carbon dioxide, methane, and sulfur dioxide in different strata of the atmosphere, among other parameters (Morgan et al., 2004; Maddy et al., 2008; Xiong et al., 2008; Engelen et al., 2009; Lin et al., 2013).

Furthermore, the data gathered from satellites and ground observations, are used in the parameterization of climatic models, which allows the calculation of additional climatic variables. The Global Land Data Assimilation System (GLDAS) is a good example of this kind of climatic modeling (Rodell et al., 2004; Rodell et al., 2009). It models land surface states and fluxes, using advanced land surface modeling techniques based on optimal fields (Rodell et al., 2004). Currently GLDAS includes datasets from four land surface models implemented in NASA's software LIS (Land Information System), namely Mosaic, Noah, the Community Land Model (CLM), and the Variable Infiltration Capacity (VIC), resulting in massive archive maps of up to 40 climatic parameters, water and energy flux, as well as underground temperature and moisture, with maximum depth of 1.1 m and with temporal coverage ranging from 1979-01-01 to nowadays (Kumar et al., 2006; Peters-Lidard et al., 2007). Another good example of a climatic model available is the Worldclim, one of the most used climatic datasets in ecological modeling. It comprises a set of 19 climatic variables relevant to many ecological processes, with a global coverage of 1000 m spatial resolution (Fick and Hijmans, 2017). This set of variables is a result of the averaging of climatic parameters from 1970 to 2000, modeled through the usage of general circulation models (GCM), which are suitable to model worldwide geographic variation in ecological processes that respond to spatial patterns of climatic heterogeneity. The calculation methods to produce this set of variables were implemented in R and are available through the function *biovars*, from the Package 'dismo', version 1.1-4 (Hijmans et al., 2017). In addition, Worldclim also

TABLE 4 | Databases for spatially explicit abiotic ecological data for use in community modeling.

Database	Data	Synopsis	References	Data access
Atmospheric Infra-Red Sounder (AIRS)	Greenhouse gases concentration in troposphere (CO2, CO, CH4, O3); etc.	Provides atmospheric chemical composition measurements by decomposing the infrared radiation in 2378 bands	AIRS Science team and Texeira, 2008; Morgan et al., 2004; Maddy et al., 2008; Xiong et al., 2008; Engelen et al., 2009; Lin et al., 2013	https://search.earthdata.nasa.gov
Tropical Rainfall Measuring Mission (TRMM)	Precipitation	Precipitation rate and rainfall rate. Was operational from 1997-12-01 to 2015-03-31	Wilheit et al., 1991	https://search.earthdata.nasa.gov
GPM (Global Precipitation Measurement)	Precipitation	Global observations of rain and snow. Operational from 2014-03-01 until the present	Hong et al., 2004; Huffman et al., 2007; Stocker et al., 2018	https://search.earthdata.nasa.gov
MODIS (Moderate Resolution Imaging Spectroradiometer)	Land surface temperature; Vegetation idexes (NDVI, EVI, LAI); Primary production; Evapotransiration; Ocean chlorophyll; etc…	Produces a huge list of high precision environmental products, with high temporal resolution, that are validated with field data	Cohen et al., 2003; Didan, 2015; Friedl and Sulla-Menashe, 2015; Giglio et al., 2015; Running et al., 2017; Savtchenko et al., 2004; Turner et al., 2006; Wan et al., 2015	https://search.earthdata.nasa.gov
SOILGRID	Bulk density; Soil granulometry; Soil classification; Cation exchange capacity; Soil organic content; pH; etc…	Models a set of soil's physical and chemical properties through the combination of soil samples data with a large set of soil covariates using machine learning techniques	Hengl et al., 2017	https://soilgrids.org
GLDAS—Global Land Data Assimilation System Version 2	Rain precipitation rate; Evapotranspiration; Root zone soil moisture; Soil moisture (in various depths); Soil temperature (in various depths); etc.	Models land surface states and fluxes using optimal fields. Includes 40 climatic parameters with temporal coverage from 1979-01-01 to present with high temporal resolution	Rodell et al., 2004; Rodell et al., 2009; Kumar et al., 2006; Peters-Lidard et al., 2007	https://search.earthdata.nasa.gov
WorldClim Version2	Annual Mean Temperature; Mean Diurnal Range; Temperature Seasonality; Temperature Annual Range; Annual Precipitation; Precipitation Seasonality; etc…	Set of 19 bioclimatic variables averaging of climatic parameters from 1970 to 2000, modeled through general circulation models (GCM).	Fick and Hijmans, 2017	http://worldclim.org/version2
WorldClim 1.4 downscaled (CMIP5) data	The same as WorldClim Version2 projected to the future	Future projections for the same WorldClim 19 bioclimatic variables for two periods, 2050 (average for 2041–2060) and 2070 (average for 2061–2080), based Intergovernmental Panel on Climate Change (IPCC)	Stocker, 2014	https://www.worldclim.org/cmip5v1

provides future projections for the same set of 19 climatic variables for two periods, 2050 (average for 2041–2060) and 2070 (average for 2061–2080), based on the set of models used in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) for the four scenarios of greenhouse gases concentration (Stocker et al., 2014). These future projections provided by Worldclim have the advantage of being bias corrected, using the current climate Worldclim data as base line, making the three sets of variables compatible. In addition, the AIRS, TRMM, GPM, and GLDAS products are available in NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC), which is part of the Earthdata platform, specialized in processing and distribution of climatic data.

Given the huge amount of climatic and environmental data available to the global landscape, microbial ecologists are now using those same analytical tools used in traditional macroecological studies. This allows them to select the most important drivers in predicting microbial diversity distribution patterns and to predict the structure of microbial communities across the globe, thereby accessing cause and effect associations. In these efforts, machine learning approaches, especially classification or regression Random Forest analysis and structural equation modeling (SEM) should be highlighted (Breiman, 2001; Grace, 2006). Specifically, Random Forest analysis constitutes specific algorithms of statistical methods of classification and regression trees (CARTs) that use binary division or regression, respectively, to form a set of trees where the importance of each predictor is inferred by decreased prediction accuracy through the random permutation of the values of these predictors (Liaw and Wiener, 2002; Wei et al., 2010). SEM routines are then used in microbial ecology studies coupled with Random Forest in order to reveal the relation between those '*a priori*' selected abiotic drivers and the target-variable in question, such as the Shannon Index, used as a proxy for microbial diversity (Delgado-Baquerizo et al., 2016). Therefore, SEM is a valuable alternative when the objective is to detail the specific relationships between multiple predictors and the modeled variable, separating them as individual pathways in the network of relationships that characterizes natural systems (Delgado-Baquerizo et al., 2017).

Incorporating Biotic Interactions in Modeling Microbial Communities

Another important issue in macroecological modeling is the inclusion of biotic interactions as predictor variables. There is an increasing evidence that species interactions improve the explanatory and predictive power of species distribution models, based on environmental variables for macroorganisms (Araújo and Luoto, 2007). Usually the inclusion of biotic interactions in species distribution models is based on previous biological knowledge of the studied species and uses a limited number of species/taxa per model, while considering their geographical distribution (Araújo and Luoto, 2007; Wisz et al., 2013; de Araújo et al., 2014). These models are usually based on species distribution models and use a maximal entropy approach—e.g., Maxent for modeling (Phillips and Dudík, 2008). However, there are also integrative modeling approaches that incorporate co-occurrence patterns into species distribution models (Pollock et al., 2014). Other modeling techniques use machine learning approaches, such as neural networks, which do not make assumptions related to species occurrence probabilities and linear relationships among environmental and biological variables, and so provide more realistic assemblage models (Harris, 2015).

Studies with micro-organisms have also suggested that including biotic interactions is necessary to build suitable predictive models (Larsen et al., 2012). However, despite their importance, these interactions can be elusive to detect, and unraveling the interactions network in microbial communities is an ongoing challenge (Faust and Raes, 2012). Biotic interactions can be inferred to some extent from co-occurrence networks (*Describing Community Structure With Co-ccurrence Networks*), but the increase of computational capacity and the development of accurate machine learning and network modeling methods has made possible to explore new approaches to statistically assess biotic interactions from large abundance datasets, such as Bayesian networks (BNs) and Genetic Algorithms (GA). The BNs are graphical models consisting of a set of variables (represented as nodes in the network) and directed arcs that describe the sets of conditional dependencies between these variables, as well as the joint probability distribution among then (Pearl, 2014; see also **Figure in Box 2**). The variables set in BNs may be both abiotic factors as well as biotic interactions, and the model can be calibrated with the same input abundance matrices generated by taxonomic annotation methods (*Taxonomic Profiling and Exploratory Analyses in Microbial Macroecology*). Additional columns representing abiotic aspects of each sampling site can be added to the abundance matrix to represent the abiotic environment experienced by a specific microorganism. This approach allows the creation of species distribution models by taking into account both biotic and abiotic aspects simultaneously in a model across large geographical scales (Staniczenko et al., 2017). These models can be further used to predict the change in the abundance of an organism when any other node (either an abiotic aspect or another species abundance) changes in the environment. A few microbial studies have already used a BN approach to study, e.g., the bacterial diversity in gut microbiota for patients with psoriatic arthritis (Scher et al., 2014) and the gut microbiota in HIV positive patients (Vázquez-Castellanos et al., 2015). Similarly, in macroecology, a few studies have used the BN approach, e.g., for range prediction of California grassland community (Staniczenko et al., 2017) and assessment of threat status of pacific walrus population in Russian and Alaskan waters at four different time periods (scenarios) throughout the twenty-first century (Jay et al., 2011).

Similarly, the use of predictive models based on the genetic algorithm (GA) method holds great potential to infer microbial interactions but has not been explored by microbiologists so far, to the best of our knowledge. The GA is an approach to solve problems inspired by the process of natural selection. Genetic programming (GP) is a particular type of GA that can be used to generate computational artifacts, such as computer programs, mathematical models, and logical models, that help to explain an observed data (Koza, 1992). The GP approach usually starts from a population of programs (algorithms) that show random levels of success in solving a task (in this case, describing the significant biotic interactions observed in a microbiome dataset). The fittest programs, that is, those best describing the data, are selected for reproduction and may undergo some "mutation" according to predefined parameters. This process is repeated over several generations in an analogy to natural selection, and the final generations are expected to show a population of much fitter programs than the initial ones. This procedure is essentially a heuristic search technique that looks for an optimal or at least suitable program among the space of all programs available. Since the construction of the models is totally guided by data, without the need of *a priori* hypotheses, the greatest potential of this technique is to generate hypotheses about the relationship between micro-organisms, as well as between micro-organisms and environment, that can be assessed by other approaches (such as BNs, dynamical modeling or common correlative statistics, described above). Applications of GP include designing electrical circuits (Koza et al., 2000), reverse engineering biochemical reactions (Sugimoto et al., 2005) and describing epidemiological relationships (Veiga et al., 2018).

Another promising approach to resolve microbial interactions is the use of dynamical models (Widder et al., 2016), which can bridge the gap between fundamental ecological knowledge and empirical interactions between taxa, by relying on explicit and mechanistically sound hypotheses. For such purpose, several modelling approaches are available (reviewed by Song et al., 2014 and by Succurro and Ebenhöh, 2018), each presenting its own set of assumptions concerning biotic and abiotic components of community. The most widespread approach is assuming direct biotic interactions among taxa and representing these interactions by using the generalized Lotka-Volterra model (gLV). This is a particular case of the population dynamic model, which can then serve to investigate concepts related to community dynamics such as

BOX 2 | Bayesian Networks: Advantages and Drawbacks.

Bayesian networks show several advantages that support their recent application in complex fields, such as: 1) network modularity, being able to integrate multiple ecosystem components (Chen and Pollino, 2012; Nojavan et al., 2014; Nojavan et al., 2017; Uusitalo, 2007), such as in management decisions field, where it is possible to integrate several sub-models as social, ecological and economic aspects (Chen and Pollino, 2012); 2) the capability of dealing with complex and nonlinear systems (Uusitalo, 2007; Aguilera et al., 2011; Phan et al., 2016; Beuzen et al., 2018); 3) possibility of incorporating expert knowledge (Uusitalo, 2007; Aguilera et al., 2011; Alameddine et al., 2011; Death et al., 2015; Phan et al., 2016), through blacklists (i.e., unrealistic relationships that are not allowed in the model) and whitelist (i.e., relationships already known in the literature); 4) being able to use a small number of samples (Uusitalo, 2007; Phan et al., 2016) 5) simplicity and little difficulty in interpreting outputs, even for non-modelers (Aguilera et al., 2011; Death et al., 2015); 6) being a rather "open" approach, different from other methods, which can be considered complicated "black-box" approaches (Chen and Pollino, 2012); 7) being able to handle high dimensional systems with the proper number of samples (Aguilera et al., 2011); 8) dealing with missing data through conditional probabilities or Bayes theorem (Uusitalo, 2007; Aguilera et al., 2011; Death et al., 2015), and finally 9) presenting less computational cost to analyze and compare different scenarios, such as climatic changes, by setting variables states in the model (Chen and Pollino, 2012; Death et al., 2015).

The main weakness of the BN approach is the lack of feedback possibilities in the model, due to it being directed acyclic graph (DAG; Phan et al., 2016). This can be bypassed by integrating models. The most critical drawback pointed in most studies is the discretization of continuous variables (Uusitalo, 2007; Aguilera et al., 2011; Nojavan A. et al., 2014; Death et al., 2015; Phan et al., 2016). The principal argument is that it causes an inevitable loss of information from data, linear relationships and consequently model performance (Uusitalo, 2007; Nojavan A. et al., 2017; Beuzen et al., 2018). However, using discrete values allows for better modeling of non-linear relationships between variables, as well as complex distributions such as bi- or multimodal distributions and can introduce greater robustness against error (Hartemink, 2001). As alternatives, there are models that could handle continuous data and not have mathematical restrictions, such as Mixture of Truncated Exponentials (MTE) models and the BN created for continuous variables (Qian and Miltner, 2015). However, it is hard to find simple examples and they are not easily found in any commercial software, which makes implementation difficult for non-modelers.

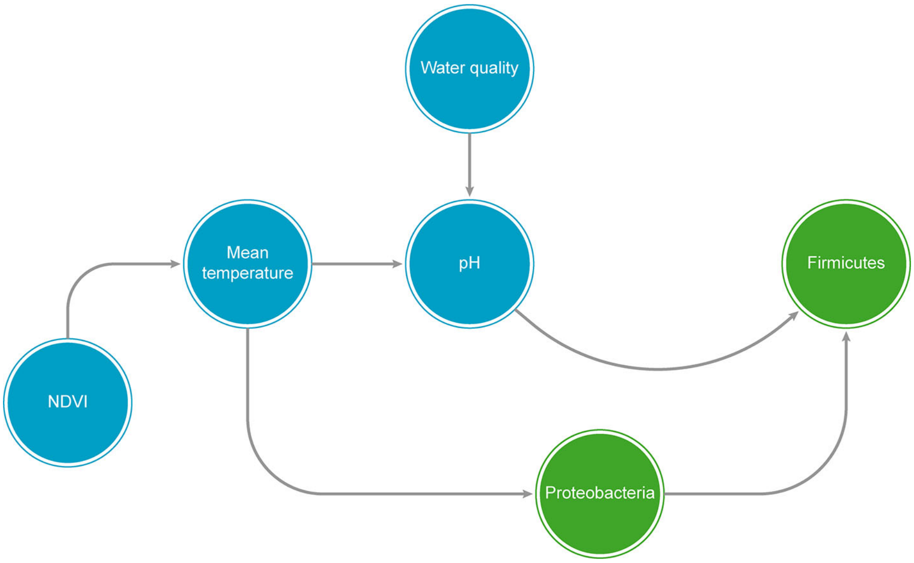


FIGURE IN BOX 2 | A graphical example of a hypothetical Bayesian Network (BN), showing both biological taxa (green circles) and predictor abiotic variables (blue circles). NDVI = Normalized difference vegetation index.

co-occurrence networks and keystone taxa (Berry and Widder, 2014; see **Box 1**). Some authors also advocate the use of metabolic-explicit dynamical models that integrate aspects of community and environmental variables, such as stoichiometry-based models and flux balance analysis (FBA; Song et al., 2014). While these approaches avoid black-box modeling and provide valuable insights into community functioning across environments, they present parameterization challenges, in gLV for instance, the number of parameters increases with the square of the number of interacting species, hindering model analysis. Future developments integrating dynamical modeling and statistical parameterization techniques are thus poised to improve the suitability of dynamical modeling approaches to exploration of microbial community interactions; meanwhile, dynamical modeling is readily available to investigate important subsystems with fewer interacting organisms.

Species Distribution Modeling for Community Prediction

The steps described in *Using Abiotic Variables to Model Microbial Communities* and *Incorporating Biotic Interactions in Modeling Microbial Communities* allow us to highlight important abiotic environmental factors as well as biotic interactions necessary to model our focal microbial communities. Although few of the techniques presented, such as BNs, can model community composition on their own, another approach largely used in macroecology for this purpose is the set of modeling tools known as species distribution modeling (SDM). The use of

SDM has been regarded as a well-established approach that can be used to overcome the lack of species spatial data, and holds great advantages for micro-organisms, a group in which the Wallacean deficit (i.e., the lack of information about species distribution) tends to be high. The SDM techniques are generally based on the concept of species ecological niches, which is the set of biotic and abiotic conditions that allows a species to persist indefinitely in a location (Soberón, 2007). Evidence so far suggests that biotic interactions should have a larger importance at smaller scales (but see Gotelli et al., 2010 and Araújo and Rozenfeld, 2013), while abiotic conditions, such as climate, should have a larger influence at larger spatial scales (McGill, 2010). Based on this, macroecologists have used the set of climatic conditions where a macroorganisms lives to estimate its potential geographic distribution. Whereas this is largely efficient for macroorganisms, more empirical evidence is necessary to evaluate these premises for micro-organisms.

Two sets of approaches can be used for SDMs: the mechanistic and correlative species distribution modeling (**Figure 5**). Mechanistic SDMs use information obtained from *ex-situ* experiments that indicate the environmental conditions that a species can tolerate (e.g., maximum and minimum temperature). This information on physiological tolerances can then be used to map areas that are environmentally suitable for the species, which can be transformed into presence/absence information (Kearney and Porter, 2009; **Figure 5**). The lack of experimental information indicating species tolerance have limited the use of mechanistic approaches; however, in areas where experimental data is abundant, such as agricultural science, mechanistic models have been used to predict potential areas for determined crop varieties (e.g., Nabout et al., 2012). This approach can be potentially useful for microbial macroecology, since these organisms can be easily manipulated *ex-situ,* because of their small, short life span and large population sizes (Jessup et al., 2004).

The correlative approach, on the other hand, uses statistical associations between acknowledged species occurrences and environmental conditions to estimate the Grinellian Niche (**Figure 5**). The type of statistical model used for this approach is then chosen upon the type of occurrence data available: continuous (abundance data), binary (i.e., presence/absence data) or presence-only data (usually the latter, since abundance information is not always available and real absence data is challenging to confirm). Presence-only models of species distributions are largely used for macroecological studies, with several algorithms available, from

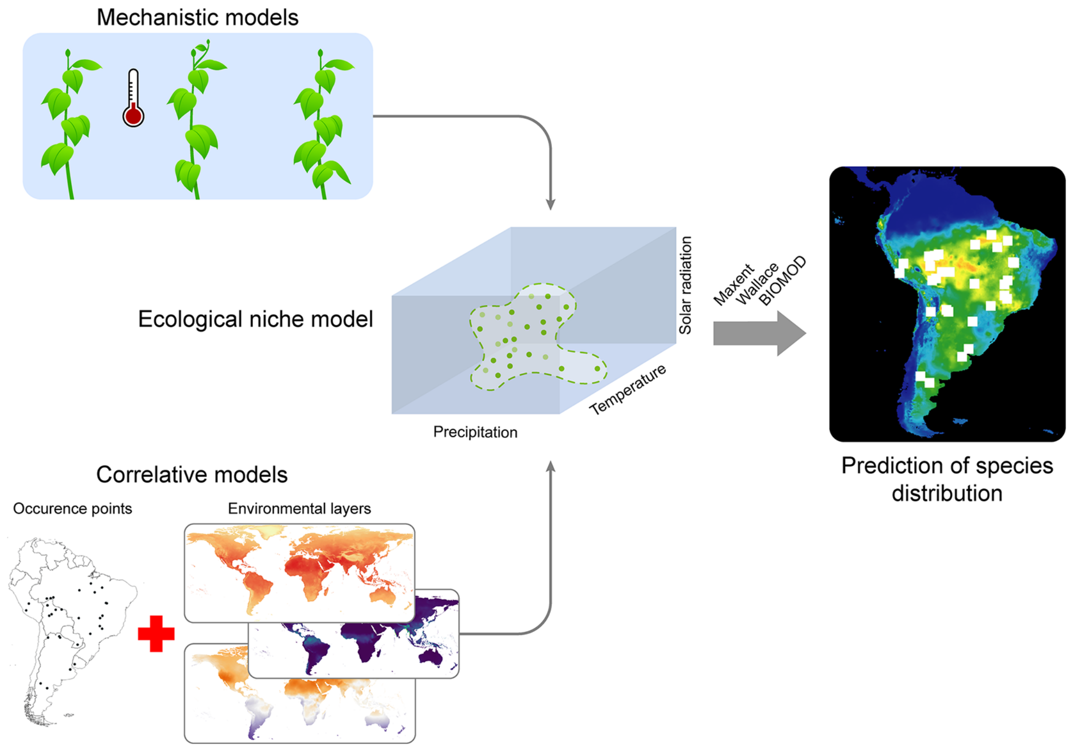
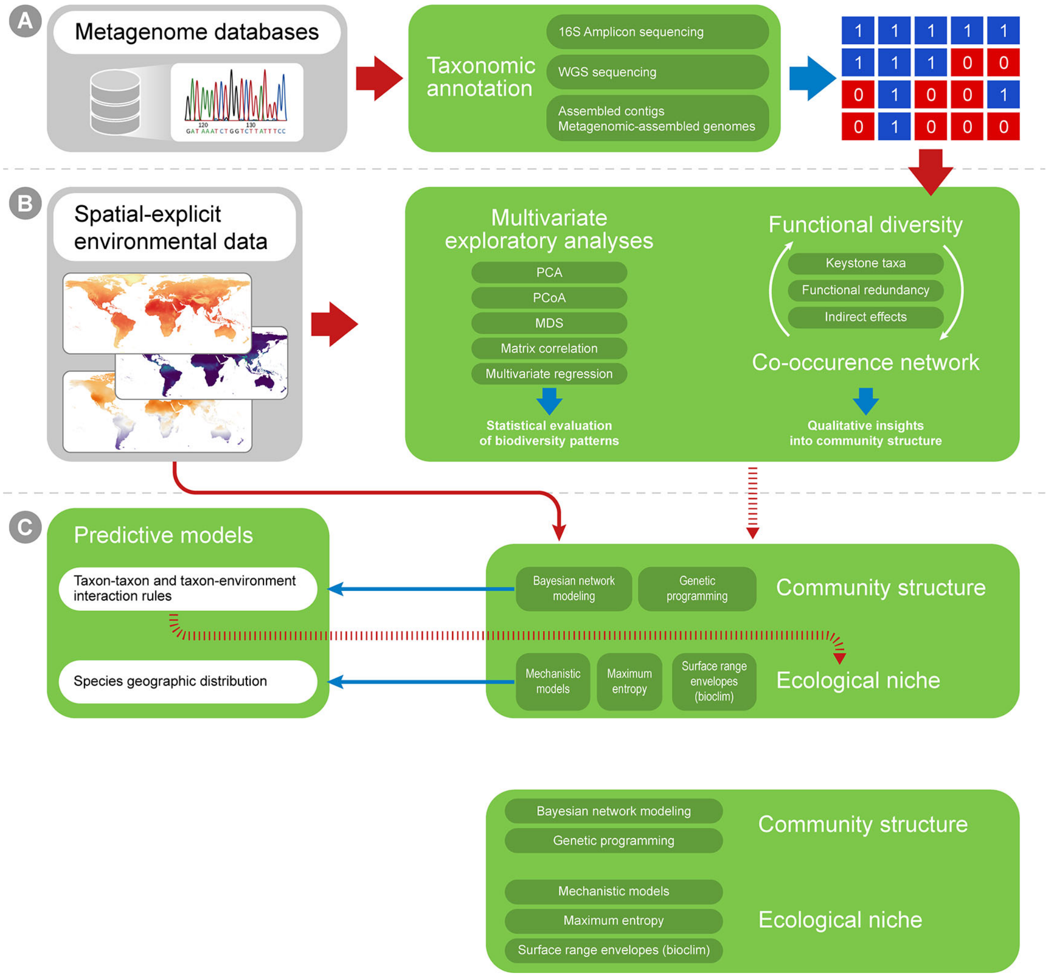


FIGURE 5 | A workflow on techniques for species distribution modelling. Ecological niches can be modeled both by using mechanistic models (upper left figure, representing temperature laboratory manipulative experiments on plants) or by using correlative models (lower left figure, representing the use of spatial-explicit environmental data combined with the knowledge about occurrence points for the species). The ecological niche is then calibrated on an n-hyperdimensional volume defined by all predictor variables used in the study (only three dimensions are shown in the cube to the center). Green points indicate known occurrence for the species projected into the environmental space; dashed green lines represent the ecological niche inferred from those points. The inferred ecological niche can then be projected into geographical space, which consists on the geographical areas having environmental conditions within those inferred to be the species' niche (are highlighted as suitable areas for the species in the map). Since the niche is statistically calibrated, i.e., as a statistical relation between predictor environmental variables and presence-absence response variables, the final map shows a gradient of environmental suitability for the species across the space.



1	1	1	1	1
1	1	1	0	0
0	1	0	0	1
0	1	0	0	0

FIGURE 6 | A methodological framework to investigate the macroecology of micro-organisms. The framework shows methods related to **(A)** gathering taxonomic data on environmental samples, **(B)** exploring the data with exploratory analyses as well as statistical tests (e.g., correlation and regression analyses), and **(C)** using the data to create predictive models about the presence/absence of species across different environments. Solid red arrows indicates input and output data that is used as input for analyses, and blue arrows indicate the output of these analyses. Dashed red arrows indicate data that can yield indirect insights for an analysis (although they are not commonly used as direct data input for the method). Grey boxes indicate external information sources and green boxes indicate the methodological approaches reviewed in this manuscript. Dark green boxes within green boxes indicate the specific techniques used in each approach. White boxes indicate the final outputs for the macroecological approach, i.e., models explaining how environment and biotic interactions affect species presence-absence and ultimately community composition. **(A)** Data from metagenomic databases can be annotated taxonomically to yield presence-absence or abundance matrixes for several ecosystems. **(B)** Spatial-explicit environmental data can be incorporated into exploratory analyses (such as PCA and MDS) as well as correlation analyses (such as regression and Mantel test) to investigate micro-organisms diversity patterns on global scales. Functional diversity can also be investigated on macroecological scales (both directly inferred from sequence reads or from the taxonomic annotation of samples). Co-occurrence networks are commonly used in microbiology studies and can yield interesting insights when different groups of samples are compared across an environmental gradient. The understanding of functional diversity and functional redundancy can be coupled with co-occurrence networks to infer the existence of keystone taxa, as well as the extent of direct and indirect effects throughout a network, and then describe the community structure and ecosystem functioning. Such structure can then be compared across macroecological scales (e.g., analyzing how the importance of specific taxa as keystone taxa varies across different environments). **(C)** Spatial-explicit environmental data can also be incorporated into models to understand community structure (such as Bayesian network modeling and genetic programming) as well as models to calibrate ecological niche (such as mechanistic and correlative niche models). These models can incorporate insights from analyses shown in **(B)**. Similarly, insights on biotic interactions, derived from community structure models, can be incorporated into ecological niche models (which commonly only use abiotic environmental variables as predictors). The final predictive models will allow microbiologists to understand interaction rules structuring microbial communities, predict the present of important taxa in different environments and infer microbial community composition across the globe.

simple ones, such as the BIOCLIM, up to more complex models based on machine learning techniques - e.g, Random Forest and MAXENT (Elith and Leathwick, 2009). While some authors claim that some algorithms have a better performance than others, the current view is that the choice of the algorithm also depends on the context in which SDMs are applied (see Peterson et al., 2010). Despite the known importance of abiotic conditions to determine large-scale species distributions, one must consider also current and historical movement limitations, such as geographical barriers, dispersal capacity and biogeographical history (Barve et al., 2011). However, it is still necessary to identify whether and how movement limitations are important to model microbial distributions, because of their overall high dispersal capacity.

Several computational tools can be used to apply SDMs, many of them freely available, open source, and collaborative (e.g., Naimi and Araújo, 2016; Kass et al., 2018). Microbiology can benefit from these methods in many research lines, since SDMs have been used not only to predict individual species distribution, but also species richness and composition (e.g., Guisan and Rahbek, 2011), species potential invasive areas (e.g., Smolik et al., 2010), as well as to understand niche evolution and speciation patterns (e.g., Silva et al., 2014; Silva et al., 2016a), and past species dynamics (e.g., Nogués-Bravo, 2009); and to model geographical range responses to climate change (e.g., Pecl et al., 2017). Specifically, SDMs present an important method to understand how species geographic range may respond to climate change. However, because of high microbial adaptation capacity, it may be a methodological challenge for microbiologists to incorporate evolution when trying to model species distribution into other time periods (Ofori et al., 2017).

CONCLUSION

The vast amount of microbial community data available represents an exciting prospect for advancing the field of microbial macroecology. In this review, we outlined the main questions in macroecology, community ecology and addressed how microbial ecologists can address them with bioinformatics, statistical and modeling tools. We covered fundamental aspects of biodiversity, reviewed classical approaches used in microbial ecology in a macroecological context, and highlighted the existing caveats and solutions to implement ecological modeling of microbial communities, which is a crucial research area for both the theoretical and practical aspects of macroecology. These approaches can serve as a general framework for microbial macroecology, addressing the two-part focus of macroecology: describing community patterns (and their drivers) at large scales and predicting community composition across the globe (**Figure 6**). The framework we present here consists of 1) gathering biological data to generate an abundance matrix, and environmental data to generate an environmental matrix; 2) exploring the associations between biological and environmental data at macroecological scales, using exploratory and network approaches; 3) incorporating insights from the previous step into modeling tools for community prediction.

The main difficulties for this research avenue are the theoretical implications derived from the biology of micro-organisms, such as higher dispersal capacity, higher evolutionary rate and the putative environmental drivers of community composition. New studies are necessary to address which environmental factors are relevant for modeling microbial distribution and to define whether the high dispersal capacity of micro-organisms makes this aspect uninformative for biogeographic patterns (i.e. the classic statement of "Everything is everywhere"). Also to evaluate whether the adaptive potential of micro-organisms is indeed high enough to violate the usual assumption of niche conservatism applied to ecological modeling. The insights from these future studies will have great impact on microbial ecological model interpretation. We predict that the development of modeling methods and approaches used in microbial macroecology, an exciting and flourishing field, will significantly contribute to the unification of microbial ecology and macroecology.

AUTHOR CONTRIBUTIONS

RM designed all figures. AT-S built Table 1. AC built Tables 2 and 3. EM built Table 3. All authors listed co-authored and proof-read the manuscript. All authors approved the final manuscript.

FUNDING

This work was primarily supported by the Serrapilheira Institute (grant number Serra-1709-17818). Further support was provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. AT-S thanks CAPES (88887.301758/2018-00). PMM thanks to PROPESQ-UFBA (11268). FC was supported by APOSTD/2018/186 Post-doctoral fellowship from Generalitat Valenciana. BD was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004. PG was supported by FAPESP (2017/08406-7) and CNPq. AC was supported by CNPq (132261/2018-9). RM was supported by CAPES (88882.460937/2019-01). AA was supported by São Paulo Research Foundation (FAPESP) grant number 2016/14277-2, and CAPES. RV was supported by CNPq, grant number 151157/2018-9.

ACKNOWLEDGMENTS

We would like to thank Gustavo Athayde for improving figures design and Noriko Cassman for language and grammar editing of the manuscript text.

REFERENCES

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environ. Model. Sofftw.* 26, 1376–1388. doi: 10.1016/j.envsoft.2011.06.004

AIRS Science team and Texeira, J. (2008). Monthly CO2 in the free troposphere (AIRS-only) 2.5 degrees x 2 degrees V005 [Data set]. *Goddard Earth Sci. Data Inf. Serv. Cent. (GES DISC)*. doi: 10.5067/Aqua/AIRS/DATA336

Alameddine, I., Cha, Y., and Reckhow, K. H. (2011). An evaluation of automated structure learning with bayesian networks: an application to estuarine chlorophyll dynamics. *Environ. Model. Soft.* 26, 163–172. doi: 10.1016/j.envsoft.2010.08.007

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amend, A. S., Oliver, T. A., Amaral-Zettler, L. A., Boetius, A., Fuhrman, J. A., Horner-Devine, M. C., et al. (2013). Macroecological patterns of marine bacteria on a global scale. *J. Biogeogr.* 40, 800–811. doi: 10.1111/jbi.12034

Anderson, M. J., Ellingsen, K. E., and McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.* 9, 683–693. doi: 10.1111/j.1461-0248.2006.00926.x

Araújo, M. B., and Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Glob. Ecol. Biogeogr.* 16, 743–753. doi: 10.1111/j.1466-8238.2007.00359.x

Araújo, M. B., and Pearson, R. G. (2005). Equilibrium of species' distributions with climate. *Ecography* 28, 693–695. doi: 10.1111/j.2005.0906-7590.04253.x

Araújo, M. B., and Rozenfeld, A. (2013). The geographic scaling of biotic interactions. *Ecography* 6, no–no. doi: 10.1111/j.1600-0587.2013.00643.x

Araújo, M. B., Rozenfeld, A., Rahbek, C., and Marquet, P. A. (2011). Using species co-occurrence networks to assess the impacts of climate change. *Ecography* 34, 897–908. doi: 10.1111/j.1600-0587.2011.06919.x

Astorga, A., Oksanen, J., Luoto, M., Soininen, J., Virtanen, R., and Muotka, T. (2012). Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules? *Glob. Ecol. Biogeogr.* 21, 365–375. doi: 10.1111/j.1466-8238.2011.00681.x

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* 31, 3322–3329. doi: 10.1093/bioinformatics/btv364

Banerjee, S., Kirkby, C. A., Schmutter, D., Bissett, A., Kirkegaard, J. A., and Richardson, A. E. (2016). Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biol. Biochem.* 97, 188–198. doi: 10.1016/j.soilbio.2016.03.017

Banerjee, S., Schlaeppi, K., and van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nat. Rev. Microbiol.* 16, 567–576. doi: 10.1038/s41579-018-0024-1

Barberán, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 6, 343–351. doi: 10.1038/ismej.2011.119

Barberán, A., Casamayor, E. O., and Fierer, N. (2014). The microbial contribution to macroecology. *Front. Microbiol.* 5, 203. doi: 10.3389/fmicb.2014.00203

Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., Dunn, R. R., et al. (2015). Continental-scale distributions of dust-associated bacteria and fungi. *Proc. Natl. Acad. Sci. U. S. A.* 112, 5756–5761. doi: 10.1073/pnas.1420815112

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Modell.* 222, 1810–1819. doi: 10.1016/j.ecolmodel.2011.02.011

Bastida, F., García, C., Fierer, N., Eldridge, D. J., Bowker, M. A., Abades, S., et al. (2019). Global ecological predictors of the soil priming effect. *Nat. Commun.* 10, 3481. doi: 10.1038/s41467-019-11472-7

Bell, T. (2010). Experimental tests of the bacterial distance–decay relationship. *ISME J.* 4, 1357. doi: 10.1038/ismej.2010.77

Berry, D., and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* 5, 219. doi: 10.3389/fmicb.2014.00219

Beuzen, T., Marshall, L., and Splinter, K. D. (2018). A comparison of methods for discretizing continuous variables in Bayesian Networks. *Environ. Model. Software* 108, 61–66.

Blaser, M. J., Cardon, Z. G., Cho, M. K., Dangl, J. L., Donohue, T. J., Green, J. L., et al. (2016). Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *MBio* 7, 1–16. doi: 10.1128/mBio.00714-16

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bond-Lamberty, B., Bolton, H., Fansler, S., Heredia-Langner, A., Liu, C., McCue, L. A., et al. (2016). Soil respiration and bacterial structure and function after 17 years of a reciprocal soil transplant experiment. *PloS One* 11, e0150599. doi: 10.1371/journal.pone.0150599

Bovallius, A., Roffey, R., and Henningson, E. (1980). Long-range transmission of bacteria. *Ann. N. Y. Acad. Sci.* 353, 186–200. doi: 10.1111/j.1749-6632.1980.tb18922.x

Bowman, J. S., and Ducklow, H. W. (2015). Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula. *PloS One* 10, e0135868. doi: 10.1371/journal.pone.0135868

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

Breiman, L. (2001). Random forest. *Mach. Learn.* 45, 5–32. doi: 10.17849/insm-47-01-31-39.1

Brown, S. M., Chen, H., Hao, Y., Laungani, B. P., Ali, T. A., Dong, C., et al. (2019). MGS-Fast: Metagenomic shotgun data fast annotation using microbial gene catalogs. *GigaScience* 8 (4), 1–9. doi: 10.1093/gigascience/giz020

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

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

Cardona, C., Weisenhorn, P., Henry, C., and Gilbert, J. A. (2016). Network-based metabolic analysis and microbial community modeling. *Curr. Opin. Microbiol.* 31, 124–131. doi: 10.1016/j.mib.2016.03.008

Casanoves, F., Pla, L., Di Rienzo, J. A., and Díaz, S. (2011). FDiversity: a software package for the integrated analysis of functional diversity. *Methods Ecol. Evol.* 2, 233–237. doi: 10.1111/j.2041-210X.2010.00082.x

Chase, J. M., and Leibold, M. A. (2002). Spatial scale dictates the productivity-biodiversity relationship. *Nature* 416, 427–430. doi: 10.1038/416427a

Chen, S. H., and Pollino, C. A. (2012). Good practice in Bayesian network modelling. *Environ. Model. Softw.* 37, 134–145. doi: 10.1016/j.envsoft.2012.03.012

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Austral Ecol.* 18, 117–143. doi: 10.1111/j.1442-9993.1993.tb00438.x

Cohen, W. B., Maiersperger, T. K., Yang, Z., Gower, S. T., Turner, D. P., Ritts, W. D., et al. (2003). Comparisons of land cover and LAI estimates derived from ETM+ and MODIS for four sites in North America: a quality assessment of 2000/2001 provisional MODIS products. *Remote Sens. Environ.* 88, 233–255. doi: 10.1016/j.rse.2003.06.006

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Colwell, R. K., and Rangel, T. F. (2009). Hutchinson's duality: the once and future niche. *Proc. Natl. Acad. Sci. U. S. A.* 106 (Suppl 2), 19651–19658.

Comte, J., Lovejoy, C., Crevecoeur, S., and Vincent, W. F. (2016). Co-occurrence patterns in aquatic bacterial communities across changing permafrost landscapes. *Biogeosciences* 13, 175–190. doi: 10.5194/bg-13-175-2016

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science* 350, 663–666. doi: 10.1126/science.aad2602

Crawley, M. J., and Harral, J. E. (2001). Scale dependence in plant biodiversity. *Science* 291, 864–868. doi: 10.1126/science.291.5505.864

de Araújo, C. B., Marcondes-Machado, L. O., and Costa, G. C. (2014). The importance of biotic interactions in species distribution models: a test of the Eltonian noise hypothesis using parrots. *J. Biogeogr.* 41, 513–523. doi: 10.1111/jbi.12234

Death, R. G., Death, F., Stubbington, R., Joy, M. K., and van den Belt, M. (2015). How good are Bayesian belief networks for environmental management? A test with data from an agricultural river catchment. *Freshw. Biol.* 60, 2297–2309. doi: 10.1111/fwb.12655

Debastiani, V. J., and Pillar, V. D. (2012). SYNCSA—R tool for analysis of metacommunities based on functional traits and phylogeny of the community components. *Bioinformatics* 28, 2067–2068. doi: 10.1093/bioinformatics/bts325

Delgado-Baquerizo, M., Maestre, F. T., Reich, P. B., Trivedi, P., Osanai, Y., Liu, Y. R., et al. (2016). Carbon content and climate variability drive global soil bacterial diversity patterns. *Ecol. Monograph.* 86 (3), 373–390. doi: 10.1002/ecm.1216/suppinfo

Delgado-Baquerizo, M., Eldridge, D. J., Maestre, F. T., Karunaratne, S. B., Trivedi, P., Reich, P. B., et al. (2017). Climate legacies drive global soil carbon stocks in terrestrial ecosystems. *Sci. Adv.* 3, e1602008. doi: 10.1126/sciadv.1602008

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., et al. (2018). A global atlas of the dominant bacteria found in soil. *Science* 359, 320–325. doi: 10.1126/science.aap9516Z

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

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., and Mouquet, N. (2010). Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. *Ecol. Lett.* 13, 1030–1040. doi: 10.1111/j.1461-0248.2010.01493.x

Dıaz, S., and Cabido, M. (2001). Vive la difference: plant functional diversity matters to ecosystem processes. *Trends Ecol. Evol.* 16, 646–655. doi: 10.1016/S0169-5347(01)02283-2

Didan, K. (2015). MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006 [Data set]. *NASA EOSDIS LP DAAC*. doi: 10.5067/MODIS/MOD13A3.006

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., et al. (2008). Functional metagenomic profiling of nine biomes. *Nature* 452, 629–632. doi: 10.1038/nature06810

Drenovsky, R. E., Steenwerth, K. L., Jackson, L. E., and Scow, K. M. (2010). Land use and climatic factors structure regional patterns in soil microbial communities. *Glob. Ecol. Biogeogr.* 19, 27–39. doi: 10.1111/j.1466-8238.2009.00486.x

Duarte, L., da, S., Carlucci, M. B., and Pillar, V. D. (2009). Macroecological analyses reveal historical factors influencing seed dispersal strategies in Brazilian Araucaria forests. *Glob. Ecol. Biogeogr.* 18, 314–326. doi: 10.1111/j.1466-8238.2009.00448.x

Elith, J., and Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697. doi: 10.1146/annurev.ecolsys.110308.120159

Engelen, R. J., Serrar, S., and Chevallier, F. (2009). Four-dimensional data assimilation of atmospheric CO 2 using AIRS observations. *J. Geophys. Res.* 114, 631. doi: 10.1029/2008JD010739

Fan, K., Weisenhorn, P., Gilbert, J. A., Shi, Y., Bai, Y., and Chu, H. (2018). Soil pH correlates with the co-occurrence and assemblage process of diazotrophic communities in rhizosphere and bulk soils of wheat fields. *Soil Biol. Biochem.* 121, 185–192. doi: 10.1016/j.soilbio.2018.03.017

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* 31, 3172–3180. doi: 10.1093/bioinformatics/btv349

Faure, D., and Joly, D. (2016). "9 - Functional Ecology and Population Genomics," in *Insight on Environmental Genomics*. Eds. D. Faure and D. Joly (Amsterdam, Netherlands: Elsevier), 93–102.

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

Faust, K., and Raes, J. (2016). CoNet app: inference of biological association networks using Cytoscape. *F1000Res* 5, 1519. doi: 10.12688/f1000research.9050.2

Fenchel, T., and Finlay, B. J. (2004). The Ubiquity of Small Species: Patterns of Local and Global Diversity. *Bioscience* 54, 777–784. doi: 10.1641/0006-3568(2004)054[0777:tuossp]2.0.co

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

Fierer, N., and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 103, 626–631. doi: 10.1073/pnas.0507535103

Fierer, N., McCain, C. M., Meir, P., Zimmermann, M., Rapp, J. M., Silman, M. R., et al. (2011). Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology* 92, 797–804. doi: 10.1890/10-1170.1

Finlay, B. J., and Clarke, K. J. (1999). Ubiquitous dispersal of microbial species. *Nature* 400, 828–828. doi: 10.1038/23616

Friedl, M., and Sulla-Menashe, D. (2015). MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006 [Data set]. *NASA EOSDIS L. Process. DAAC*. doi: 10.5067/MODIS/MCD12C1.006

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

Fu, H., Zhong, J., Fang, S., Hu, J., Guo, C., Lou, Q., et al. (2017). Scale-dependent changes in the functional diversity of macrophytes in subtropical freshwater lakes in south China. *Sci. Rep.* 7, 8294. doi: 10.1038/s41598-017-08844-8

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., et al. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 105, 7774–7778. doi: 10.1073/pnas.0803070105

Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C., and Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J.* 12, 2470–2478. doi: 10.1038/s41396-018-0158-1

Giglio, L., Justice, C., Boschetti, L., and Roy, D. (2015). MCD64A1 MODIS/Terra+Aqua Burned Area Monthly L3 Global 500m SIN Grid V006 [Data set]. *NASA EOSDIS L. Process. DAAC*. doi: 10.5067/MODIS/MCD64A1.006

Gotelli, N. J., Graves, G. R., and Rahbek, C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proc. Natl. Acad. Sci. U. S. A.* 107, 5030–5035. doi: 10.1073/pnas.0914089107

Grace, J. B. (2006). *Structural equation modeling natural systems* (Cambridge, UK: Cambridge University Press).

Grenié, M., Denelle, P., Tucker, C. M., Munoz, F., and Violle, C. (2017). funrar: An R package to characterize functional rarity. *Divers. Distrib.* 23, 1365–1371. doi: 10.1111/ddi.12629

Guimarães, P. R. Jr., Pires, M. M., Jordano, P., Bascompte, J., and Thompson, J. N. (2017). Indirect effects drive coevolution in mutualistic networks. *Nature* 550, 511–514. doi: 10.1038/nature24273

Guimera, R., and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895–900. doi: 10.1038/nature03288

Guisan, A., and Rahbek, C. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.* 38, 1433–1444. doi: 10.1111/j.1365-2699.2011.02550.x

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., et al. (2013). Big data and the future of ecology. *Front. Ecol. Environ.* 11, 156–162. doi: 10.1890/120103

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., and Martiny, J. B. H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* 10, 497–506. doi: 10.1038/nrmicro2795

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.* 6, 465–473. doi: 10.1111/2041-210X.12332

Hartemink, A. J. (2001). Principled computational methods for the validation discovery of genetic regulatory networks, (Doctoral dissertation, Massachusetts Institute of Technology). Available at: https://dspace.mit.edu/handle/1721.1/8699?show=full [Accessed August 19, 2019].

Hartman, K., van der Heijden, M. G. A., Wittwer, R. A., Banerjee, S., Walser, J.-C., and Schlaeppi, K. (2018). Cropping practices manipulate abundance patterns of root and soil microbiome members paving the way to smart farming. *Microbiome* 6, 14. doi: 10.1186/s40168-017-0389-9

Hatzenpichler, R. (2012). Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. *Appl. Environ. Microbiol.* 78, 7501–7510. doi: 10.1128/AEM.01960-12

Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J., and Classen, A. T. (2017). Consistently inconsistent drivers of microbial diversity and abundance at macroecological scales. *Ecology* 98, 1757–1763. doi: 10.1002/ecy.1829

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 12, e0169748. doi: 10.1371/journal.pone.0169748

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., and Hijmans, M. R. J. (2017). Package ‘dismo.’. *Circles* 9, 1–68. doi: 10.1002/joc.5086

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *Am. Nat.* 163, 192–211. doi: 282400/381004

Hong, Y., Hsu, K.-L., Sorooshian, S., and Gao, X. (2004). Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural Network Cloud Classification System. *J. Appl. Meteorol.* 43, 1834–1853. doi: 10.1175/JAM2173.1

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc. Natl. Acad. Sci. U. S. A.* 106 Suppl 2, 19659–19665. doi: 10.1073/pnas.0905137106

Horner-Devine, M. C., Lage, M., Hughes, J. B., and Bohannan, B. J. M. (2004). A taxa-area relationship for bacteria. *Nature* 432, 750–753. doi: 10.1038/nature03073

Horner-Devine, M. C., Silver, J. M., Leibold, M. A., Bohannan, B. J. M., Colwell, R. K., Fuhrman, J. A., et al. (2007). A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* 88, 1345–1353. doi: 10.1890/06-0286

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., and Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.* 46, 523–549. doi: 10.1146/annurev-ecolsys-112414-054400

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeorol.* 8, 38–55. doi: 10.1175/JHM560.1

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016a). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi: 10.1038/nmicrobiol.2016.48

Hug, L. A., Thomas, B. C., Sharon, I., Brown, C. T., Sharma, R., Hettich, R. L., et al. (2016b). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* 18, 159–173. doi: 10.1111/1462-2920.12930

Hugenholtz, P., and Tyson, G. W. (2008). Microbiology: metagenomics. *Nature* 455, 481–483. doi: 10.1038/455481a

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., et al. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol. Direct* 13, 6. doi: 10.1186/s13062-018-0208-7

Hutchins, D. A., Mulholland, M. R., and Fu, F. (2009). Nutrient Cycles and Marine Microbes in a CO_2-Enriched Ocean. *Oceanography* 22, 128–145. doi: 10.5670/oceanog.2009.103

Jackson, M. A., Bonder, M. J., Kuncheva, Z., Zierer, J., Fu, J., Kurilshikov, A., et al. (2018). Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ* 6, e4303. doi: 10.7717/peerj.4303

Jarzyna, M. A., and Jetz, W. (2018). Taxonomic and functional diversity change is scale dependent. *Nat. Commun.* 9, 2565. doi: 10.1038/s41467-018-04889-z

Jay, C. V., Marcot, B. G., and Douglas, D. C. (2011). Projected status of the Pacific walrus (Odobenus rosmarus divergens) in the twenty-first century. *Polar Biol.* 34, 1065–1084. doi: 10.1007/s00300-011-0967-4

Jessup, C. M., Kassen, R., Forde, S. E., Kerr, B., Buckling, A., Rainey, P. B., et al. (2004). Big questions, small worlds: microbial model systems in ecology. *Trends Ecol. Evol.* 19, 189–197. doi: 10.1016/j.tree.2004.01.008

Jiao, S., Liu, Z., Lin, Y., Yang, J., Chen, W., and Wei, G. (2016). Bacterial communities in oil contaminated soils: Biogeography and co-occurrence patterns. *Soil Biol. Biochem.* 98, 64–73. doi: 10.1016/j.soilbio.2016.04.005

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* 88, 2427–2439. doi: 10.1890/06-1736.1

Kass, J. M., Vilela, B., Aiello-Lammens, M. E., Muscarella, R., Merow, C., and Anderson, R. P. (2018). Wallace: a flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods Ecol. Evol.* 9, 1151–1156. doi: 10.1111/2041-210X.12945

Kearney, M., and Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecol. Lett.* 12, 334–350. doi: 10.1111/j.1461-0248.2008.01277.x

Keller, C. K., White, T. M., O'brien, R., and Smith, J. L. (2006). Soil CO2 dynamics and fluxes as affected by tree harvest in an experimental sand ecosystem. *J. Geophys. Res.: Biogeosci.* 111, (G3). doi: 10.1029/2005jg000157

Kerr, J. T., Kharouba, H. M., and Currie, D. J. (2007). The macroecological contribution to global change solutions. *Science* 316, 1581–1584. doi: 10.1126/science.1133267

Koslicki, D., and Falush, D. (2016). MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. *mSystems* 1, 1–18. doi: 10.1128/mSystems.00020-16

Koza, J. R., Bennett, F. H., III, Andre, D., and Keane, M. A. (2000). Synthesis of topology and sizing of analog electrical circuits by means of genetic programming. *Comput. Methods Appl. Mech. Eng.* 186, 459–482. doi: 10.1109/4235.687879

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (Cambridge, Massachusetts, USA: MIT Press).

Kultima, J. R., Coelho, L. P., Forslund, K., Huerta-Cepas, J., Li, S. S., Driessen, M., et al. (2016). MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32, 2520–2523. doi: 10.1093/bioinformatics/btw183

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., et al. (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environ. Model. Softw.* 21, 1402–1415. doi: 10.1016/j.envsoft.2005.07.004

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

Laliberté, E., and Legendre, P. (2010). A distance-based framework for measuring functional diversity from multiple traits. *Ecology* 91, 299–305.

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

Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* 9, 621–625. doi: 10.1038/nmeth.1975

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75, 5111–5120. doi: 10.1128/AEM.00335-09

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends Microbiol.* 25, 217–228. doi: 10.1016/j.tim.2016.11.008

Legendre, P., and Legendre, L. F. J. (2012). *Numerical Ecology*. (Elsevier Amsterdam, Netherlands).

Legendre, P., Borcard, D., and Peres-Neto, P. R. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.* 75, 435–450. doi: 10.1890/05-0549

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., et al. (2011a). The European Nucleotide Archive. *Nucleic Acids Res.* 39, D28–D31. doi: 10.1093/nar/gkq967

Leinonen, R., Sugawara, H., Shumway, M. International Nucleotide Sequence Database Collaboration (2011b). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology* 73, 1943–1967. doi: 10.2307/1941447

Li, D., Zhan, M., Liu, H., Liao, Y., and Liao, G. (2017). A Robust Translational Motion Compensation Method for ISAR Imaging Based on Keystone Transform and Fractional Fourier Transform Under Low SNR Environment. *IEEE Trans. Aerosp. Electron. Syst.* 53, 2140–2156. doi: 10.1109/TAES.2017.2683599

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2 (3), 18–22.

Lima-Mendez, G., Faust, K., Henry, N., and Decelle, J. (2015). Determinants of community structure in the global plankton interactome. *Science* 348 (6237), 1262073.

Lin, H.-H., and Liao, Y.-C. (2016). Accurate binning of metagenomic contigs *via* automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* 6, 24175. doi: 10.1038/srep24175

Lin, H., Yu, B., Chen, Z., Hu, Y., Huang, Y., Wu, J., et al. (2013). A geospatial web portal for sharing and analyzing greenhouse gas data derived from satellite remote sensing images. *Front. Earth Sci.* 7, 295–309. doi: 10.1007/s11707-013-0365-z

Lomolino, M. V. (2001). Elevation gradients of species-density: historical and prospective views. *Glob. Ecol. Biogeogr.* 10, 3–13. doi: 10.1046/j.1466-822x.2001.00229.x

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi: 10.1126/science.aaf4507

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., et al. (2018). Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* 2, 936–943. doi: 10.1038/s41559-018-0519-1

Lupatini, M., Suleiman, A. K. A., Jacques, R. J. S., Antoniolli, Z. I., de Siqueira Ferreira, A., Kuramae, E. E., et al. (2014). Network topology reveals high connectance levels and few key microbial genera within soils. *Front. Environ. Sci. Eng. China* 2, 343. doi: 10.3389/fenvs.2014.00010

Mace, G. M., Norris, K., and Fitter, A. H. (2012). Biodiversity and ecosystem services: a multilayered relationship. *Trends Ecol. Evol.* 27, 19–26. doi: 10.1016/j.tree.2011.08.006

Maddy, E. S., Barnet, C. D., Goldberg, M., Sweeney, C., and Liu, X. (2008). CO2 retrievals from the Atmospheric Infrared Sounder: Methodology and validation. *J. Geophys. Res. D: Atmos.* 113, (D11). doi: 10.1029/2007jd009402

Marasco, R., Rolli, E., Fusi, M., Michoud, G., and Daffonchio, D. (2018). Grapevine rootstocks shape underground bacterial microbiome and networking but not potential functionality. *Microbiome* 6, 3. doi: 10.1186/s40168-017-0391-2

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4, 102–112. doi: 10.1038/nrmicro1341

May, R. M. (1972). Will a large complex system be stable? *Nature* 238, 413–414. doi: 10.1038/238413a0

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

McGill, B. J., and Nekola, J. C. (2010). Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos* 119, 591–603. doi: 10.1111/j.1600-0706.2009.17771.x

McGill, B. (2003). Strong and weak tests of macroecological theory. *Oikos* 102, 679–685. doi: 10.1034/j.1600-0706.2003.12617.x

McGill, B. J. (2010). Ecology. Matters of scale. *Science* 328, 575–576. doi: 10.1126/science.1188528

Mendes, L. W., Mendes, R., Raaijmakers, J. M., and Tsai, S. M. (2018). Breeding for soil-borne pathogen resistance impacts active rhizosphere microbiome of common bean. *ISME J.* 12, 3038–3042. doi: 10.1038/s41396-018-0234-6

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf.* 9, 386. doi: 10.1186/1471-2105-9-386

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* 46, D726–D735. doi: 10.1093/nar/gkx967

Morgan, C. G., Allen, M., Liang, M. C., Shia, R. L., Blake, G. A., and Yung, Y. L. (2004). Isotopic fractionation of nitrous oxide in the stratosphere: Comparison between model and observations. *J. Geophys. Res. D: Atmos.* 109, (D4) doi: 10.1029/2003jd003402

Mouchet, M. A., Villéger, S., Mason, N. W. H., and Mouillot, D. (2010). Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Funct. Ecol.* 24, 867–876. doi: 10.1111/j.1365-2435.2010.01695.x

Nabout, J. C., Caetano, J. M., Ferreira, R. B., Teixeira, I. R., and de Freitas Alves, S. M. (2012). Using Correlative, Mechanistic and Hybrid Niche Models to Predict the Productivity and Impact of Global Climate Change on Maize Crop in Brazil. *Natureza Conservação* 10, 177–183. doi: 10.4322/natcon.2012.034

Naimi, B., and Araújo, M. B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography* 39, 368–375. doi: 10.1111/ecog.01881

Nelson, M. B., Martiny, A. C., and Martiny, J. B. H. (2016). Global biogeography of microbial nitrogen-cycling traits in soil. *Proc. Natl. Acad. Sci. U. S. A.* 113, 8033–8040. doi: 10.1073/pnas.1601070113

Nogués-Bravo, D. (2009). Predicting the past distribution of species climatic niches. *Glob. Ecol. Biogeogr.* 18, 521–531. doi: 10.1111/j.1466-8238.2009.00476.x

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Noguez, A. M., Arita, H. T., Escalante, A. E., Forney, L. J., García-Oliva, F., and Souza, V. (2005). Microbial macroecology: highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Glob. Ecol. Biogeogr.* 14, 241–248. doi: 10.1111/j.1466-822X.2005.00156.x

Nojavan, A. F., Qian, S. S., Paerl, H. W., Reckhow, K. H., and Albright, E. A. (2014). A study of anthropogenic and climatic disturbance of the New River Estuary using a Bayesian belief network. *Mar. Pollut. Bull.* 83, 107–115. doi: 10.1016/j.marpolbul.2014.04.011

Nojavan, A. F., Qian, S. S., and Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environ. Model. Softw.* 87, 64–71. doi: 10.1016/j.envsoft.2016.10.007

Nottingham, A. T., Fierer, N., Turner, B. L., Whitaker, J., Ostle, N. J., McNamara, N. P., et al. (2018). Microbes follow Humboldt: temperature drives plant and soil microbial diversity patterns from the Amazon to the Andes. *Ecology* 99, 2455–2466. doi: 10.1002/ecy.2482

Ofori, B. Y., Stow, A. J., Baumgartner, J. B., and Beaumont, L. J. (2017). Influence of adaptive capacity on the outcome of climate change vulnerability assessment. *Sci. Rep.* 7, 12979. doi: 10.1038/s41598-017-13245-y

Ohgushi, T. (2005). Indirect Interaction Webs: Herbivore-Induced Effects Through Trait Change in Plants. *Annu. Rev. Ecol. Evol. Syst.* 36, 81–105. doi: 10.1146/annurev.ecolsys.36.091704.175523

Oliver, T. H., Heard, M. S., Isaac, N. J. B., Roy, D. B., Procter, D., Eigenbrod, F., et al. (2015). Biodiversity and Resilience of Ecosystem Functions. *Trends Ecol. Evol.* 30, 673–684. doi: 10.1016/j.tree.2015.08.009

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16, 236. doi: 10.1186/s12864-015-1419-2

Paine, R. T. (1966). Food Web Complexity and Species Diversity. *Am. Nat.* 100, 65–75. doi: 282400/282400

Paine, R. T. (1969). The Pisaster-Tegula interaction: prey patches, predator food preference, and intertidal community structure. *Ecology* 50, 950–961. doi: 10.2307/1936888

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. (Elsevier, Amsterdam, Netherlands).

Peay, K. G., Garbelotto, M., and Bruns, T. D. (2010). Evidence of dispersal limitation in soil microorganisms: isolation reduces species richness on mycorrhizal tree islands. *Ecology* 91, 3631–3640. doi: 10.1890/09-2237.1

Pecl, G. T., Araújo, M. B., Bell, J. D., Blanchard, J., Bonebrake, T. C., Chen, I.-C., et al. (2017). Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* 355, 1–9. doi: 10.1126/science.aai9214

Petchey, O. L., and Gaston, K. J. (2002). Functional diversity (FD), species richness and community composition. *Ecol. Lett.* 5, 402–411. doi: 10.1046/j.1461-0248.2002.00339.x

Petchey, O. L., and Gaston, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecol. Lett.* 9, 741–758. doi: 10.1111/j.1461-0248.2006.00924.x

Petchey, O. L., Hector, A., and Gaston, K. J. (2004). How do different measures of functional diversity perform? *Ecology* 85, 847–857. doi: 10.1890/03-0226

Peters-Lidard, C. D., Houser, P. R., Tian, Y., Kumar, S. V., Geiger, J., Olden, S., et al. (2007). High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innov. Syst. Software Eng.* 3, 157–165. doi: 10.1007/s11334-007-0028-x

Peterson, A. T., Knapp, S., Guralnick, R., Soberón, J., and Holder, M. T. (2010). The big questions for biodiversity informatics. *Syst. Biodivers.* 8, 159–168. doi: 10.1080/14772001003739369

Phan, T. D., Smart, J. C. R., Capon, S. J., Hadwen, W. L., and Sahin, O. (2016). Applications of Bayesian belief networks in water resource management: A systematic review. *Environ. Model. Softw.* 85, 98–111. doi: 10.1016/j.envsoft.2016.08.006

Phillips, S. J., and Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31, 161–175. doi: 10.1111/j.0906-7590.2008.5203.x

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., et al. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.* 5, 397–406. doi: 10.1111/2041-210X.12180

Poudel, R., Jumpponen, A., Schlatter, D. C., Paulitz, T. C., Gardener, B. B. M., Kinkel, L. L., et al. (2016). Microbiome Networks: A Systems Framework for Identifying Candidate Microbial Assemblages for Disease Management. *Phytopathology* 106, 1083–1096. doi: 10.1094/PHYTO-02-16-0058-FI

Qian, S. S., and Miltner, R. J. (2015). A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA. *Environ. Model. Softw.* 69, 14–22. doi: 10.1016/j.envsoft.2015.03.001

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

Rahbek, C. (2005). The role of spatial scale and the perception of large-scale species-richness patterns. *Ecol. Lett.* 8, 224–239. doi: 10.1111/j.1461-0248.2004.00701.x

Ramirez, K. S., Döring, M., Eisenhauer, N., Gardi, C., Ladau, J., Leff, J. W., et al. (2015). Toward a global platform for linking soil biodiversity data. *Front. Ecol. Evol.* 3, 2189. doi: 10.3389/fevo.2015.00091

Ramirez, K. S., Knight, C. G., de Hollander, M., Brearley, F. Q., Constantinides, B., Cotton, A., et al. (2018). Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nat. Microbiol.* 3, 189–196. doi: 10.1038/s41564-017-0062-x

Ren, Z., Wang, F., Qu, X., Elser, J. J., Liu, Y., and Chu, L. (2017). Taxonomic and Functional Differences between Microbial Communities in Qinghai Lake and Its Input Streams. *Front. Microbiol.* 8, 2319. doi: 10.3389/fmicb.2017.02319

Ricotta, C., de Bello, F., Moretti, M., Caccianiga, M., Cerabolini, B. E. L., and Pavoine, S. (2016). Measuring the functional redundancy of biological communities: a quantitative guide. *Methods Ecol. Evol.* 7, 1386–1395. doi: 10.1111/2041-210X.12604

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi: 10.1146/annurev.genet.38.072902.091216

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., et al. (2004). The Global Land Data Assimilation System. *Bull. Am. Meteorol. Soc* 85, 381–394. doi: 10.1175/BAMS-85-3-381

Rodell, M., Velicogna, I., and Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. *Nature* 460, 999–1002. doi: 10.1038/nature08238

Running, S., Mu, Q., and Zhao, M.. (2017). MOD16A3 MODIS/Terra Net Evapotranspiration Yearly L4 Global 500m SIN Grid V006 [Data set]. *NASA EOSDIS L. Process. DAAC.* doi: 10.5067/MODIS/MOD16A3.006

Safi, K., Cianciaruso, M. V., Loyola, R. D., Brito, D., Armour-Marshall, K., and Diniz-Filho, J. A. F. (2011). Understanding global patterns of mammalian functional and phylogenetic diversity. *Philos. Trans. R. Soc Lond. B Biol. Sci.* 366, 2536–2544. doi: 10.1098/rstb.2011.0024

Savtchenko, A., Ouzounov, D., Ahmad, S., Acker, J., Leptoukh, G., Koziana, J., et al. (2004). Terra and Aqua MODIS products available from NASA GES DAAC. *Adv. Sp. Res.* 34, 710–714. doi: 10.1016/j.asr.2004.03.012

Scher, J. U., Bretz, W. A., and Abramson, S. B. (2014). Periodontal disease and subgingival microbiota as contributors for rheumatoid arthritis pathogenesis: modifiable risk factors? *Curr. Opin. Rheumatol.* 26, 424–429. doi: 10.1097/BOR.0000000000000076

Schleuter, D., Daufresne, M., Massol, F., and Argillier, C. (2010). A user's guide to functional diversity indices. *Ecol. Monogr.* 80, 469–484. doi: 10.1890/08-2225.1

Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371–4378. doi: 10.1128/jb.173.14.4371-4378.1991

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

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. doi: 10.1038/nmeth.4458

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J. M., Herrmann, M., et al. (2018). Macroecology to Unite All Life, Large and Small. *Trends Ecol. Evol.* 33, 731–744. doi: 10.1016/j.tree.2018.08.005

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Silva, D. P., Vilela, B., De Marco, P. Jr., and Nemésio, A. (2014). Using ecological niche models and niche analyses to understand speciation patterns: the case of sister neotropical orchid bees. *PloS One* 9, e113246. doi: 10.1371/journal.pone.0113246

Silva, D. P., Vilela, B., Buzatto, B. A., Moczek, A. P., and Hortal, J. (2016a). Contextualized niche shifts upon independent invasions by the dung beetle Onthophagus taurus. *Biol. Invasions* 18, 3137–3148. doi: 10.1007/s10530-016-1204-4

Silva, G. G. Z., Green, K. T., Dutilh, B. E., and Edwards, R. A. (2016b). SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* 32, 354–361. doi: 10.1093/bioinformatics/btv584

Smolik, M. G., Dullinger, S., Essl, F., Kleinbauer, I., Leitner, M., Peterseil, J., et al. (2010). Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. *J. Biogeogr.* 37, 411–422. doi: 10.1111/j.1365-2699.2009.02227.x

Snyder, L. A. S., Loman, N., Pallen, M. J., and Penn, C. W. (2009). Next-generation sequencing–the promise and perils of charting the great microbial unknown. *Microb. Ecol.* 57, 1–3. doi: 10.1007/s00248-008-9465-9

Soberón, J., and Nakamura, M. (2009). Niches and distributional areas: concepts, methods, and assumptions. *Proc. Natl. Acad. Sci. U. S. A.* 106 (Suppl 2), 19644–19650. doi: 10.1073/pnas.0901637106

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* 10, 1115–1123. doi: 10.1111/j.1461-0248.2007.01107.x

Soberón, J. M. (2010). Niche and area of distribution modeling: a population ecology perspective. *Ecography* 33, 159–167. doi: 10.1111/j.1600-0587.2009.06074.x

Soininen, J. (2012). Macroecology of unicellular organisms–patterns and processes. *Environ. Microbiol. Rep.* 4, 10–22. doi: 10.1111/j.1758-2229.2011.00308.x

Song, H.-S., Cannon, W. R., Beliaev, A. S., and Konopka, A. (2014). Mathematical modeling of microbial community dynamics: a methodological review. *Processes* 2, 711–752. doi: 10.3390/pr2040711

Speth, D. R., In 't Zandt, M. H., Guerrero-Cruz, S., Dutilh, B. E., and Jetten, M. S. M. (2016). Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. *Nat. Commun.* 7, 11172. doi: 10.1038/ncomms11172

Staniczenko, P. P. A., Sivasubramaniam, P., Suttle, K. B., and Pearson, R. G. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecol. Lett.* 20, 693–707. doi: 10.1111/ele.12770

Stocker, T. F., Dahe, Q., and Plattner, G.-K. (2014). *Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (United Kingdom: Cambridge University Press).

Stocker, T. (2014). Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change. (United Kingdom: Cambridge University Press).

Stocker, E. F., Alquaied, F., Bilanow, S., Ji, Y., and Jones, L. (2018). TRMM Version 8 Reprocessing Improvements and Incorporation into the GPM Data Suite. *J. Atmos. Ocean. Technol.* 35, 1181–1199. doi: 10.1175/JTECH-D-17-0166.1

Succurro, A., and Ebenhöh, O. (2018). Review and perspective on mathematical modeling of microbial ecosystems. *Biochem. Soc Trans.* 46, 403–412. doi: 10.1042/BST20170265

Sugimoto, M., Kikuchi, S., and Tomita, M. (2005). Reverse engineering of biochemical equations from time-course data by means of genetic programming. *Biosystems* 80, 155–164. doi: 10.1111/j.1461-0248.2004.00701.x

Taudiere, A., and Violle, C. (2016). cati: an R package using functional traits to detect and quantify multi-level community assembly processes. *Ecography* 39, 699–708. doi: 10.1111/ecog.01433

Turner, D. P., Ritts, W. D., Cohen, W. B., Gower, S. T., Running, S. W., Zhao, M., et al . (2006). Evaluation of MODIS NPP and GPP products across multiple biomes. *Remote Sens. Environ.* 102, 282–292. doi: 10.1016/j.rse.2006.02.017

Ugarte, A., Vicedomini, R., Bernardes, J., and Carbone, A. (2018). A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome* 6 (1), 1–27. doi: 10.1186/s40168-018-0532-2

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecol. Modell.* 203, 312–318. doi: 10.1016/j.ecolmodel.2006.11.033

Vázquez-Castellanos, J. F., Serrano-Villar, S., Latorre, A., Artacho, A., Ferrús, M. L., Madrid, N., et al. (2015). Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. *Mucosal Immunol.* 8, 760–772. doi: 10.1038/mi.2014.107

Veiga, R. V., Barbosa, H. J. C., Bernardino, H. S., Freitas, J. M., Feitosa, C. A., Matos, S. M. A., et al. (2018). Multiobjective grammar-based genetic programming applied to the study of asthma and allergy epidemiology. *BMC Bioinf.* 19, 245. doi: 10.1186/s12859-018-2233-z

Vieira-Silva, S., Falony, G., Darzi, Y., Lima-Mendez, G., Garcia Yunta, R., Okuda, S., et al. (2016). Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* 1, 16088. doi: 10.1038/nmicrobiol.2016.88

von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., and Dutilh, B. E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *bioRxiv* 530188, 1–14. doi: 10.1101/530188

Wan, Z., Hook, S., and Hulley, G.. (2015). MOD11B3 MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 6km SIN Grid V006 [Data set]. *NASA EOSDIS LP DAAC*. doi: 10.5067/MODIS/MOD11B3.006

Webb, C. O., Ackerly, D. D., McPeek, M. A., and Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annu. Rev. Ecol. Syst.* 33, 475–505. doi: 10.1146/annurev.ecolsys.33.010802.150448

Webb, C. O., Ackerly, D. D., and Kembel, S. W. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* 24, 2098–2100. doi: 10.1093/bioinformatics/btn358

Wei, C. L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M. J., Pitcher, C. R., et al. (2010). Global patterns and predictions of sea- floor biomass using random forests. *PloS One* 5 (12), e15323.

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

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., et al. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J.* 10, 2557–2568. doi: 10.1038/ismej.2016.45

Wilheit, T. T., Chang, A. T. C., and Chiu, L. S. (1991). Retrieval of Monthly Rainfall Indices from Microwave Radiometric Measurements Using Probability Distribution Functions. *J. Atmos. Ocean. Technol.* 8, 118–136. doi: 10.1175/1520-0426(1991)008<0118:ROMRIF>2.0.CO;2

Willig, M. R., Kaufman, D. M., and Stevens, R. D. (2003). Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis. *Annu. Rev. Ecol. Evol. Syst.* 34, 273–309. doi: 10.1146/annurev.ecolsys.34.012103.144032

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., et al. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev. Camb. Philos. Soc* 88, 15–30. doi: 10.1111/j.1469-185X.2012.00235.x

Wu, J., Shen, W., Sun, W., and Tueller, P. T. (2002). Empirical patterns of the effects of changing scale on landscape metrics. *Landsc. Ecol.* 17, 761–782. doi: 10.1023/A:1022995922992

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

Xiong, X., Barnet, C., Maddy, E., Sweeney, C., Liu, X., Zhou, L., et al. (2008). Characterization and validation of methane products from the Atmospheric Infrared Sounder (AIRS). *J. Geophys. Res.* 113, 253. doi: 10.1029/2007JG000500

Xue, P.-P., Carrillo, Y., Pino, V., Minasny, B., and McBratney, A. B. (2018). Soil Properties Drive Microbial Community Structure in a Large Scale Transect in South Eastern Australia. *Sci. Rep.* 8, 11725. doi: 10.1038/s41598-018-30005-8

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi: 10.1038/nbt.1823

Zhou, J., Deng, Y., Shen, L., Wen, C., Yan, Q., Ning, D., et al. (2016). Temperature mediates continental-scale diversity of microbes in forest soils. *Nat. Commun.* 7, 12083. doi: 10.1038/ncomms12083

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

frontiers in Genetics

METHODS
published: 23 January 2020
doi: 10.3389/fgene.2019.01366

NG-Tax 2.0: A Semantic Framework for High-Throughput Amplicon Analysis

Wasin Poncheewin[1†], *Gerben D. A. Hermes*[2†], *Jesse C. J. van Dam*[1], *Jasper J. Koehorst*[1], *Hauke Smidt*[2] *and Peter J. Schaap*[1*]

[1] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, Netherlands, [2] Laboratory of Microbiology, Wageningen University & Research, Wageningen, Netherlands

OPEN ACCESS

Edited by:
Jens Stoye,
Bielefeld University, Germany

Reviewed by:
Cuncong Zhong,
University of Kansas, United States
Sebastian Jünemann,
Bielefeld University, Germany

***Correspondence:**
Peter J. Schaap
peter.schaap@wur.nl

†These authors have contributed equally to this work

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

Received: *05 March 2019*
Accepted: *12 December 2019*
Published: *23 January 2020*

Citation:
Poncheewin W, Hermes GDA, van Dam JCJ, Koehorst JJ, Smidt H and Schaap PJ (2020) NG-Tax 2.0: A Semantic Framework for High-Throughput Amplicon Analysis. Front. Genet. 10:1366. doi: 10.3389/fgene.2019.01366

NG-Tax 2.0 is a semantic framework for FAIR high-throughput analysis and classification of marker gene amplicon sequences including bacterial and archaeal 16S ribosomal RNA (rRNA), eukaryotic 18S rRNA and ribosomal intergenic transcribed spacer sequences. It can directly use single or merged reads, paired-end reads and unmerged paired-end reads from long range fragments as input to generate *de novo* amplicon sequence variants (ASV). Using the RDF data model, ASV's can be automatically stored in a graph database as objects that link ASV sequences with the full data-wise and element-wise provenance, thereby achieving the level of interoperability required to utilize such data to its full potential. The graph database can be directly queried, allowing for comparative analyses of over thousands of samples and is connected with an interactive Rshiny toolbox for analysis and visualization of (meta) data. Additionally, NG-Tax 2.0 exports an extended BIOM 1.0 (JSON) file as starting point for further analyses by other means. The extended BIOM file contains new attribute types to include information about the command arguments used, the sequences of the ASVs formed, classification confidence scores and is backwards compatible. The performance of NG-Tax 2.0 was compared with DADA2, using the plugin in the QIIME 2 analysis pipeline. Fourteen 16S rRNA gene amplicon mock community samples were obtained from the literature and evaluated. Precision of NG-Tax 2.0 was significantly higher with an average of 0.95 vs 0.58 for QIIME2-DADA2 while recall was comparable with an average of 0.85 and 0.77, respectively. NG-Tax 2.0 is written in Java. The code, the ontology, a Galaxy platform implementation, the analysis toolbox, tutorials and example SPARQL queries are freely available at http://wurssb.gitlab.io/ngtax under the MIT License.

Keywords: operational taxonomic unit, amplicon sequence variants, taxonomic classification, FAIR, semantic web, RDF, ontology, SPARQL

INTRODUCTION

High-throughput sequencing technologies have empowered our ability to study complex environmental and host-associated microbial communities. Of these technologies, amplicon sequencing targeting marker genes is currently the most cost-effective tool to assess the microbial composition of large numbers of samples (Tringe and Rubin, 2005; Yarza et al., 2014; Stulberg et al., 2016). By using smart multiplexing techniques hundreds of samples can be sequenced at once while sequencing costs per sample are further reduced leading to immense amounts of microbial community composition data available for large scale comparisons.

High-throughput amplicon sequencing is, however, inevitably noisy. Due to PCR artefacts and low-quality base calls, a fraction of the amplicon reads will contain one or more sequence errors (error-reads), which in turn could lead to false taxonomic inferences. One strategy to reduce the number of false taxonomic inferences due to these error-reads, is to cluster amplicon reads by sequence identity in operational taxonomic units (a process called OTU-picking) at some user defined identity thresholds. To build these OTUs, centroid or seed sequence-based greedy clustering approaches are frequently used (Stackebrandt and Goebel, 1994; Konstantinidis and Tiedje, 2005; Godzik and Li, 2006; Edgar, 2010). Centroid based OTU-picking approaches however, have a number of disadvantages as they require a predefined identity threshold, while the representative centroid sequence is influenced by selection of the seed, sequence input order and the amount of amplicon sequences and PCR error present in the sample, all of which make OTU-picking by clustering sample dependent and therefore, in principle, not suitable for comparisons between different sets of samples (Callahan et al., 2016). Recent studies have shown that a *de novo* clustering approach using exact matches would yield better results (Ramiro-Garcia et al., 2016; Callahan et al., 2017). These exact match sequence clusters have been termed Amplicon Sequence Variants (ASVs), sub-OTUs or zero-radius OTUs (Tikhonov et al., 2015; Callahan et al., 2016; Edgar, 2018). The rationale is that an ASV is not a representative sequence from a cluster of similar sequences, but is directly derived from a biological entity. An ASV can be separated from error-reads on the basis of the expectation that due to the biological origin, a real sequence variant is located at a fixed position in the amplicon sequence and therefore more likely to be repeatedly observed in those samples where the particular biological variant is present. Error-reads are assumed to be present at a relatively low abundance, and because sequence errors are also positionally dispersed (Schirmer et al., 2016) they are unable to form meaningful exact match ASV clusters. In NG-Tax, an exact match OTU-picking algorithm is used to find ASV forward and reverse sequence read pairs. Likely erroneous ASVs are rejected if their read count does not exceed an experimentally defined dynamic threshold that takes the evenness of the distribution into account (Ramiro-Garcia et al., 2016). In the past the accuracy of NG-Tax has been benchmarked against QIIME (Caporaso et al., 2010), using synthetic mock communities and has been shown to outperform it (Ramiro-Garcia et al., 2016).

Unlike centroid based OTUs which work with representative sequences, ASV sequences are believed to directly descend from an existing biological entity, and the presence of this entity can therefore be validly compared across many samples (Callahan et al., 2017). Such large-scale analyses would require tracking of multiple ASVs over multiple samples and thus a high degree of interoperability. Proper data handling can be achieved through the application of the FAIR data principles which are intended to make the information Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). We adopted these principles in NG-Tax 2.0 through implementation of a semantic framework using a Linked Data format (RDF) for data serialization and handling, combined with a strictly applied ontology. In NG-Tax 2.0 ASV amplicon sequences are automatically converted into a semantic data model, ASV objects, that link ASV sequences with the full data-wise and element-wise provenance thereby achieving the level of interoperability required to utilize such data to its full potential.

NG-Tax 2.0 is a complete redesign and rewrite of the NG-Tax amplicon analysis pipeline. In NG-Tax 2.0 many of the limitations of NG-Tax have been addressed and as a result NG-Tax 2.0 has evolved into a highly automated framework for high-throughput classification and comparative analysis of marker gene amplicon sequences.

Using ten mock communities publicly available from the Mockrobiota database (Bokulich et al., 2016) and data from four staggered mocks described by (Tourlousse et al., 2017) the precision and recall of NG-Tax 2.0 was evaluated against DADA2 (Callahan et al., 2016) using the plugin in the QIIME 2 pipeline. The known relative abundance of each ASV in these mock communities enabled a precise evaluation of the tools on how they perform in predicting the number of species, their relative abundance and their taxonomic classification. The integrative power of using a semantic framework is demonstrated by performing a meta-analysis across the mock samples and multiple reference databases.

MATERIALS AND METHODS

NG-Tax 2.0

NG-Tax 2.0 is written in Java with Gradle as build system. A Galaxy web implementation (Afgan et al., 2016) is also available. A k-bounded Levenshstein distance function (Hawkins et al., 2018) was implemented in Java to measure the edit distances between amplicon sequences in OTU-picking and between ASV sequences and reference database sequences for taxonomic annotation of ASV objects. The distance function was slightly modified to account for phantom out of word frame insertion and deletions.

NG-Tax 2.0 Semantic Framework

An NG-Tax 2.0 specific expansion of the GBOL ontology (van Dam et al., 2019) was developed in Protégé (Musen and Team, 2015). Empusa (van Dam et al., 2019) was used to convert the ontology to a Java API. As a result, picked ASVs, taxonomic inferences and linked metadata can be automatically stored in a graph database and can be directly retrieved and compared through a list of (routine) SPARQL queries. A list of routine SPARQL queries is provided (**Data Sheet 1**), the output of which directly interacts with the NG-Tax 2.0 data analysis and visualization toolbox that is based on Rshiny (Chang et al., 2017). RDF (turtle) files were imported into a local GraphDB (http://graphdb.ontotext.com/) repository and queried using the SPARQL query language.

Mock Communities

Mock communities were retrieved from the Mockrobiota project (Bokulich et al., 2016). Ten demultiplexed 16S-rRNA gene mock communities were obtained (**Table 1**).

Bioinformatic Analysis

General

The mock communities were analysed using: NG-Tax 2.0 and QIIME2 using the DADA2 plugin (Hall and Beiko, 2018, p. 2). Apart from the variation in amplicon read length, all settings remained as the default. The SILVA reference database was used for the taxonomic classification (Yilmaz et al., 2014). For comparison purposes, three incremental stable versions of the database were downloaded from https://www.arb-silva.de/download/archive/being: 123, 128 and 132 (latest). Additionally, a custom 16S rRNA gene database was created *de novo* using sequences from (Hug et al., 2016; **Data Sheet 2**) as input. For comparison the description line of the sequences was converted to contain the taxonomic lineage in the SILVA format. The chimera detection process has been described by Ramiro-Garcia et al., 2016. Briefly, chimeras are detected using the following condition: if the forward and reverse read of the ASV are identical to two different ASVs in the same sample and the abundance of the matched ASVs are at least twice of the abundance, then the ASV is marked as chimeric.

TABLE 1 | Mock communities used for NG-Tax 2.0 benchmarking.

Mockrobiota #	Composition	Read length	Reference
Mock13	21 bacterial strains, evenly distributed	250/250	Kozich et al., 2013
Mock14	21 bacterial strains, evenly distributed	250/250	
Mock15	21 bacterial strains, evenly distributed	250/250	
Mock16	49 bacterial strains, 10 archaea, evenly distributed	250/250	Schirmer et al., 2015
Mock18	15 bacterial strains, evenly distributed	250/250	Tourlousse et al., 2017
Mock19	15 bacterial strains, 12 synthetic spike-in standards, evenly distributed	250/250	
Mock20	20 bacterial strains, evenly distributed	301/301	Gohl et al., 2016
Mock21	20 bacterial strains, staggered	301/301	
Mock22	20 bacterial strains, evenly distributed	301/301	
Mock23	20 bacterial strains, staggered	301/301	
SRX1868061-SRX1868064	15 bacterial strains, 12 synthetic spike-in standards, staggered	250/250	Tourlousse et al., 2017

**All mocks utilized the V4 region.*

Lookup Table

For taxonomic annotation of ASV objects, NG-Tax 2.0 creates a lookup table from reference sequences. There are two options.

When a multiple alignment file, such as the 50,000 columns long SILVA alignment is provided, NG-Tax 2.0 assumes that the sequence of the primer region is conserved in the alignment. Using a regular expression which takes care of IUPAC wildcard characters, NG-Tax 2.0 finds in each aligned sequence, primer start and stop positions, starting with the first aligned sequence and keeps on doing this until a consensus start and stop column position is obtained (defined as: the start and the stop position of the primer are found to occur in the same columns/positions a 1,000 times). It then assumes that the region of interest is in the columns between the primer columns, extracts this region, removes alignment gaps, trims the sequences to the chosen forward and reverse read length and subsequently transforms the sequences into a four-column lookup table. An example is shown in **Table 2**.

For special cases such when strain specific markers have been developed, or for studying a new species or a designed community in a closed system, NG-Tax 2.0 can also build a custom lookup table from unaligned reference sequences. For this NG-Tax 2.0 uses a regular expression representing the (degenerate) primers used in amplification to find the region of interest, taking into account a single mismatch with the exception of the most 3-prime nucleotide of the primer which must either have a perfect match or a G/T mismatch for amplification to occur. The sequence region in between the primers is subsequently used to build the lookup table as described above. To illustrate this approach a custom 16S rRNA gene database was created *de novo* using sequences from (Hug et al., 2016, **Data Sheet 2**) as input.

NG-Tax 2.0 Configuration

To use the NG-Tax 2.0 command line interface, users need to provide the paired-end amplicon reads in comma separated format (-fS), the mapping file (-mapFile), a reference database such as the SILVA database (-refdb) for creation of the look-up

TABLE 2 | Example of the look-up table.

AGGAT…	CGACA…	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Chryseobacterium	148
AGGAT…	CGACA…	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Anaplasmataceae;Wolbachia	276
GGGAT…	CGACA…	Bacteria;Cyanobacteria;Chloroplast;Corchorus_capsularis;_;_	3
GGGAT…	CGACA…	Bacteria;Cyanobacteria;Chloroplast;Isatis_tinctoria;_;_	4
GGGAT…	CGACA…	Bacteria;Cyanobacteria;Chloroplast;Aethionema_carneum;_;_	1

table, the selected forward and the reverse primer used for selection of the amplified region in the reference file (-for_p and -rev_p), the name for the output RDF file (-t), and the name for the output BIOM file (-b), and they need to specify whether the primers were already removed from the sequences or not (-primerRemoved). Various amplicon read lengths were used in the analysis: 70, 100, 140, 200 and 240 nt. Other settings were kept as the default as the following: -minPerT 0.1, -identLvl 100, -errorCorr 1 and -classifyRatio 0.8. as described in (Ramiro-Garcia et al., 2016). A full list of options can be found at http://wurssb.gitlab.io/ngtax/commandLine.html. Parameters are stored in the output file in "args" section of the extended BIOM and RDF file.

QIIME2-DADA2 Configuration

For QIIME2, the latest SILVA database for QIIME2 (version 132) was downloaded from the official QIIME2 website at https://docs.qiime2.org/2018.11/. SILVA database version 128 was downloadable through the forum page (https://forum.qiime2.org/t/silva-128-classifiers-available-for-download/3558). Silva database version 123 needed to be created manually through q2-feature-classifier tutorial https://docs.qiime2.org/2018.6/tutorials/feature-classifier/).

To analyse the data with this pipeline, we imported reads into QIIME2 as an artefact using the Casava 1.8 paired-end demultiplexed Fastq format. DADA2 (Callahan et al., 2016) was selected as the method for quality control using the following parameters: –p-trim-left-f 19 and –p-trim-left-r 20 as the length of the primer combined with various read lengths, 140, 150, 180, 200, 220 and 240, for both –p-trunc-len-f and –p-trunc-len-r. The trimming option (–p-trim-left-f and –p-trim-left-r) was used only for mock 16, 18, 19, 22 and 23. This results in a feature table, representative sequences and a statistical outcome captured during this denoising step. Next, classify-sklearn was used to classify the taxonomic lineage of the representative sequences based on the given database. Then, the classified sequences were collapsed with the feature table in order to produce an OTU table at a certain taxonomic lineage resolution based on the user input, such as 6 for genus level. Finally, the OTU table is exported into a Hierarchical Data Format (HDF5) file format which can be converted in to a tab separated values (tsv) or a JavaScript object notation (json) file format using the BIOM package (http://biom-format.org/documentation/biom_conversion.html#general-usage-examples).

Statistics

Binary Classifier

Comparison between the expected and the predicted results using the confusion matrix.

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F - score = 2 \times \frac{precision \times recall}{precision + recall}$$

where TP is the number of true positives, FN is the number of false negatives and FP is the number of false positives. A TP was defined as an exact match at genus level.

Modified Rv Coefficient

Comparison between two weighted adjacency matrices, which in this case is the microbial composition and their relative abundance. The results can be interpreted as Pearson's correlations.

$$RV_2(X, Y) = \frac{Vec\left(\widetilde{XX'}\right)' Vec\left(\widetilde{YY'}\right)}{\sqrt{Vec\left(\widetilde{XX'}\right)' Vec\left(\widetilde{XX'}\right) \times Vec\left(\widetilde{YY'}\right)' Vec\left(\widetilde{YY'}\right)}}$$

Given matrix $\widetilde{XX'} = [XX' - diag(XX')]$, where $diag(XX')$ is a matrix containing only the diagonal elements of XX' on its diagonal, and zero's elsewhere. The same definition also applied to YY'.

RESULTS

NG-Tax 2.0 is fully written in Java and can be executed from the command line, while a Galaxy toolbox implementation (Afgan et al., 2016) is also available. Using multiplexed amplicon sequences as input, NG-Tax 2.0 executes four major tasks: demultiplexing and amplicon read cleaning, generation of ASV objects (a process generally referred to as OTU-picking), denoising and taxonomic assignment. Processed samples, derived ASV sequences, taxonomic inferences and data provenance are automatically linked and serialized in an RDF-triple store format and can be exported as an extended Biom 1.0 file for compatibility reasons (**Figure 1**).

Development of the Semantic Framework

NG-Tax 2.0 uses the RDF data model to capture and store analysis results and associated data provenance as ASV objects. To ensure consistency and to have a high degree of interoperability and reusability a strictly defined ontology was created, focusing on its function as file format and as database schema. The modular design ensures that the ontology can be extended and currently consists of eight main classes (**Table 3**).

To increase human readability, ontology class names represent the underlying concept as closely as possible. Classes start with uppercase whereas properties start with lowercase. *Library* is the root of the ontology. Each *Library* contains samples according to the input mapping file and it also refers back to the metadata and the command arguments. Each sample contains ASV objects composed of the forward and reverse sequence of the particular ASV, the number of amplicon reads in the sample that have this particular forward and reverse sequence and their taxonomic annotation. The

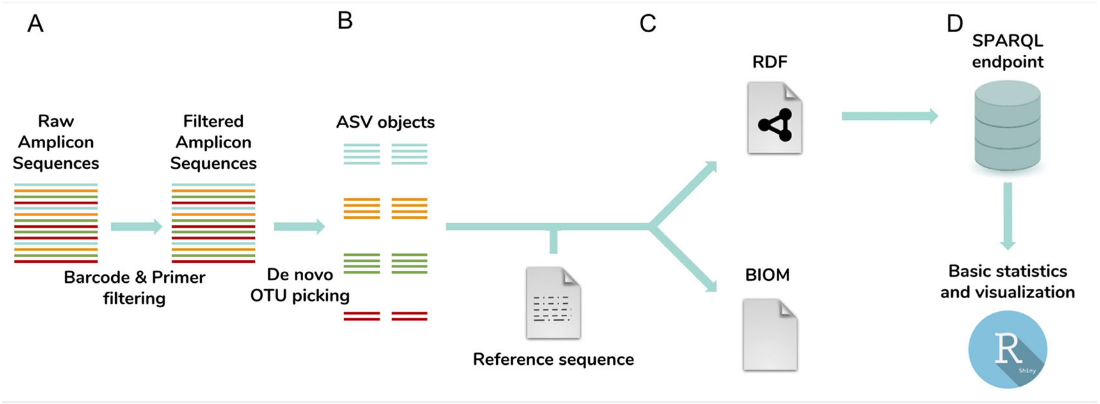


FIGURE 1 | NG-Tax 2.0 workflow. The workflow consists of four main steps: **(A)** barcode and primer filtering, **(B)** *de novo* OTU-picking of ASV sequences, artefact filtering, correction for the impact of error reads on ASV relative abundance estimates and taxonomic inference; **(C)** ASV object serialization and storage. ASV sequences, taxonomic inferences and data provenance including library and sample names and used settings are exported and stored as ASV objects in an RDF triple store graph database and optionally exported in the Biom 1.0 file format. **(D)** Downstream analysis tool box. ASV data and meta-data can be directly queried and analysed through the SPARQL endpoint. The Rshiny toolbox directly provides standard statistics and visualizations using predefined SPARQL queries.

TABLE 3 | Description of the NG-Tax 2.0 ontology main classes.

Main ontology class	Description
Library*	Description of samples in a library
Sample	Description of PCRPrimers, BarcodeSet and ASVSet
Sequence	ASVSequence: ASV forward and reverse sequences
SequenceSet*	ASVSet, RejectedAsChimera, RejectedASV BarcodeSet, PCRPrimerSet
Taxon	Taxon name and rank annotation of an ASVSet
ASVAssignment	Taxon information and related provenance
Provenance	Interlinks ProvenanceClassification, containing tool specific information with the input Library
ProvenanceClassification*	Contains confidence score of taxonomic assignment and user input command argument in the analysis

**NG-Tax 2.0 specific extensions of the gbol ontology.*

ASVAssignment class is a class where all the possible taxonomic hits of the ASV objects are stored (**Figure 2**). The NG-Tax 2.0 ontology is integrated in the Genome Biology Ontology Language available at http://gbol.life (van Dam et al., 2019).

ASV-Picking, Artefact Filtering and Correction for the Impact of Error-Reads on the Relative Abundance Estimates

NG-Tax 2.0 can handle both single and paired-end reads. In NG-Tax 2.0 paired-end reads are filtered for matching primers and barcodes but not merged and reads are subsequently processed in parallel. As the forward and reverse read may significantly differ in quality and reverse reads may require additional trimming, in NG-Tax 2.0 the forward and reverse reads are not necessarily of the same length and therefore two parameters are used (-for_read_len and -rev_read_len) to define read lengths used for ASV formation. If the -rev_read_len parameter is not set, single reads or merged forward and reverse reads can be used in the analysis.

NG-Tax 2.0 error-handling is built on the assumption that erroneous reads are more likely to be less abundant than true biological variation. In addition, it is assumed that erroneous sequences (reads with random sequencing errors and (amplified) reads systematic sequence errors) have a high degree of sequence similarity with true reads amplified from the same template sequence in the sample. To deal with such erroneous sequences NG-Tax 2.0 does not start from individual reads or read-pairs but first builds a collection of initial ASV objects from the pool of available reads. In NG-Tax 2.0 by default three (default, user defined) or more identical forward and reverse sequences will form an ASV object and the thus clustered forward and reverse sequences of this object are subsequently used as a reference sequences in the two-step error handling.

NG-Tax 2.0 first assumes that the remaining (singleton) read-pairs are unable to join an already existing ASV object because of a random sequencing error. NG-Tax 2.0 uses a k-bounded Levenshtein function and a cumulative edit distance of one nucleotide (mismatch or indel) to find a match between ASV objects and singleton read pairs. If a singleton ASV read pair shows a single mismatch (mutation or indel) with an ASV reference in either the forward or the reverse read, it is assumed this is due to a random sequence error and the singleton is joined with the particular object thereby increasing the read count of the object but not changing the original sequences linked to the object. Singletons showing more than one mismatch are considered as sample specific noise and discarded.

Secondly, due to PCR and sequence-specific errors (Shin and Park, 2016), specific amplicons may also accumulate above-average sequencing errors resulting in the formation of an erroneous ASV object. Here the assumption is that an erroneous ASV object will show a high degree of sequence similarity with an also existing true ASV object. To find erroneous ASV objects, NG-Tax 2.0 ranks ASV-objects by read counts and uses the k-bounded Levenshtein function to merge ASV objects with read-count below a set threshold, with ASV

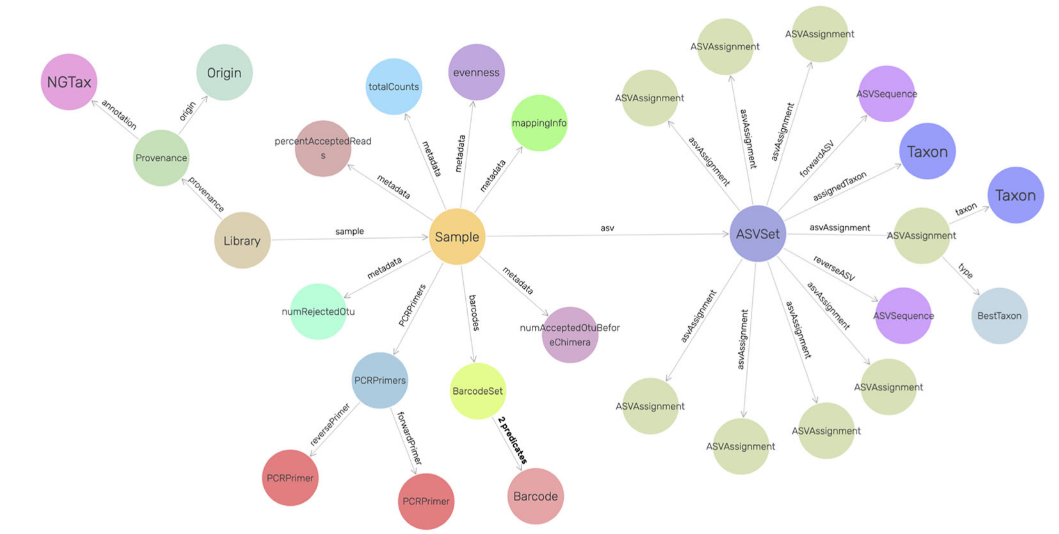


FIGURE 2 | Graphical view of the NG-Tax 2.0 data management model. Nodes are defined in the GBOL ontology. *Sample* and *ASVset* are main hubs and represent sample input and NG-Tax 2.0 processed data. Each *ASVset* represents a specific ASV object, consisting of a collection of (inter)linked descriptions of entities representing data, knowledge and associated meta data of the specific ASV. Each *ASVset* is directly linked to the *Sample* node which is used as a hub for the experimental dependencies. Each *Sample* is part of a *Library* containing information of an individual sequence and analysis run. The visualization was done in GraphDB (http://graphdb.ontotext.com/) using the visual graph interface.

objects with read counts better than the set threshold starting with the ASV object with the highest read count. If a selected ASV object below the threshold has a single mismatch (mutation or indel) with a high read-count ASV object the two ASV objects are merged. The sequences of the high read-count object are kept because they are believed to be true and the read-counts of both objects are summed. For this merging process a user defined relative abundance threshold is used and by default this is set to 0.1% of the total number of read-pairs associated with ASVs. If NG-Tax 2.0 cannot merge an ASV object with a read count below the set threshold, it will be labelled as 'provisionally rejected" but the ASV object remains in the output file for further analysis as it could be a true variation, and therefore the first 100 (default, user defined) most abundant provisionally rejected ASVs also obtain a taxonomic assignment. However, most of these flagged ASVs are likely to be sample specific noise (Faith et al., 2013). To show that provisionally rejected ASVs are likely noise we followed their fate in a closed biological system. Samples were obtained from a dietary intervention in an *in vitro* system that simulates the dynamics conditions in the human colon (**Data Sheet 3**). To show reproducibility, several replicates were taken. Because we do not delete but only label as such, sample specific provisionally rejected ASVs we can track their presence over multiple replicates and samples using SPARQL queries. The sequences of almost all provisionally rejected ASVs were only present in a single sample. The percentage of flagged as rejected ASVs that were present in at least two individual samples, ranged from 2.7 to 5.4%, which indeed suggests that the vast majority of the flagged ASVs is likely sample specific noise.

Taxonomic Assignment of ASV Objects

NG-Tax 2.0 uses reference fasta or alignment files obtained from repositories such as the ARB-SILVA database (Quast et al., 2012) for taxonomic assignments. To reduce the computational load, reference sequences are trimmed such that they include only the region matching the reads. The length of the regions of interest are defined by the length of the reads in the ASV object while the location of the amplicon primer sequences in the reference sequences are used to mark the 5'- and 3'-end of the region of interest. Subsequently, the thus reduced reference file is converted into a look-up table by clustering and counting entries that are identical in sequence and in taxonomic annotation. This look-up table is automatically re-used when different sets of samples with the same parameters are processed. Using the k-bounded Levenshtein function with an upper-bound of 50, the edit distance between each ASV read pair and entries in the reference file is measured. For each edit distance with a maximum sequence mismatch between the reference sequence and the amplicon sequence of 15%, a list of sequence entries, including frequency of occurrence in the reference database file and taxonomic annotation is generated and stored as an integral part of the particular ASV object. This list is also included in the exported extended Biom file. Following a set of rules outlined below, the classifier subsequently proposes from this list of candidates the most likely taxonomic assignment by taking into account the number of mismatches. Depending on the level of sequence identity with the reference set, by default the lowest possible taxonomic ranks proposed by NG-Tax 2.0 will be used, out of species, genus, family and order. Species will only be assigned when a perfect match is obtained with a single species.

Between 100-95% sequence identity the lowest proposed taxonomic assignment is genus, between 95-92% the level is family and below 92%, the level is order. These values are stored as attributes of the *CommandArgs* class. Note that while these rules provide a tentative taxonomic assignment based on best practices, for each ASV the full list of reference database sequences remains available and can be retrieved and compared by querying the graph database at any time through a SPARQL endpoint.

Analysis of NG-Tax 2.0 Precision and Recall Using Mock Communities

We measured NG-Tax 2.0 precision and recall using ten staggered and evenly distributed MiSeq 16S rRNA gene mock communities (**Table 1**) obtained from the Mockrobiota public repository. Then communities were analysed in parallel with the DADA2 implementation in the QIIME2 pipeline (from here on referred to as DADA2). NG-tax 2.0 and DADA2 taxonomic predictions were compared using different read lengths and three consecutive stable versions of the ARB-SILVA reference database. The reference composition of the selected mock communities is based on SILVA version 123 using a similarity threshold of 97% and 99% respectively. **Figure 3** displays a typical example showing a compositional analysis of mock21 using either NG-Tax 2.0 or DADA2. For each tool, the optimal read length was used.

The metrics used to compare and evaluate the performances of both pipelines were recall, precision and F-score. F-score is a single metric that combines both recall and precision and is used here to select an optimal read length for the analysis. When considering F-scores from both pipelines for different mock communities at different read lengths, NG-Tax 2.0 had a higher range of 0.65 to 0.97, compared to DADA2's 0.42 to 0.76, across all mock communities (**Figure 4**). Moreover, NG-Tax 2.0 revealed an optimal read length at 140 nucleotides with F-scores ranging from 0.73 to 0.97 across all the communities. In contrast, DADA2's optimal read length varied between mock communities, which suggests that the performance of this tool in

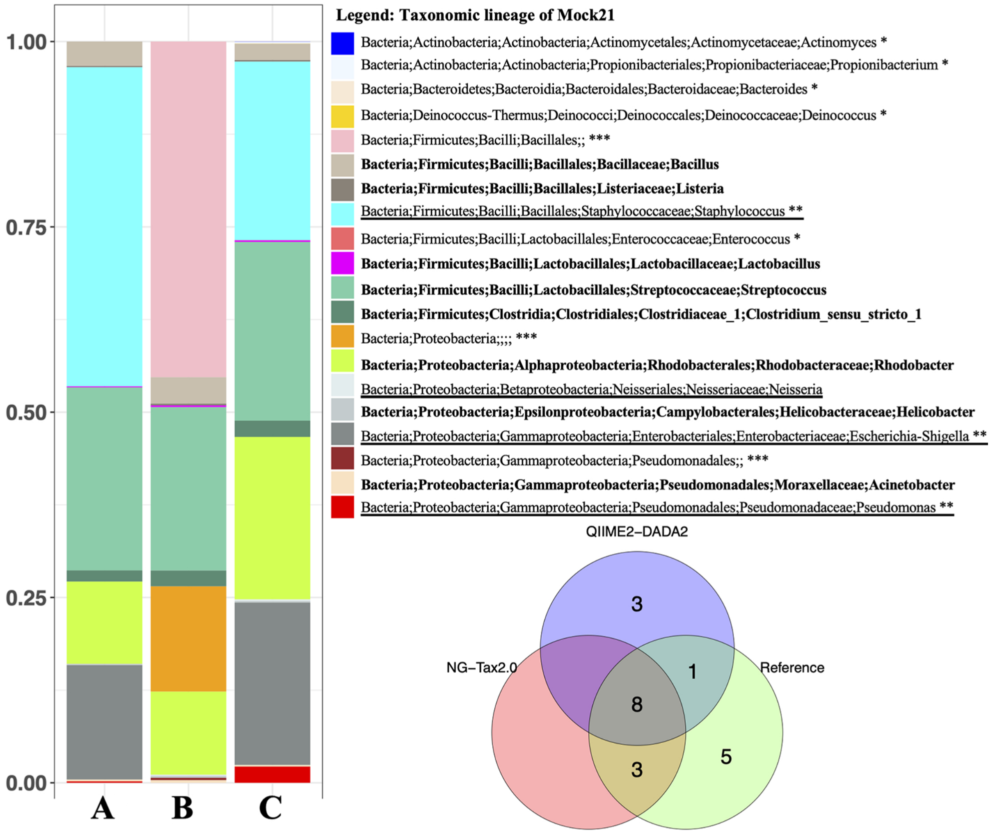


FIGURE 3 | Microbial composition of Mockrobiota mock community 21. Mock21 is a staggered mixture of 20 bacterial strains. *Left* mock21 NG-Tax 2.0 and DADA2 predictions using the ARB-Silva reference database 123 for taxonomic annotation. In this comparison, for each tool the optimal read length was used: 140nt for NG-Tax 2.0 and 220nt for DADA2. **(A)** NG-Tax 2.0. **(B)** DADA2. **(C)** Reference composition. Identical results were obtained with the reference database with 97% and 99% similarity thresholds. In bold, mock21 reference taxons correctly identified with NG-Tax 2.0 and DADA2. *Mock21 reference taxons not detected by either tool, **Mock21 reference taxons detected by NG-Tax 2.0 but not DADA2. *Underlined*, mock 21 reference taxons detected by DADA2 but not NG-Tax 2.0. ***No prediction at genus level, however correctly assigned the taxonomic lineage. *Right* Venn diagram summarizing the taxonomic annotation results.

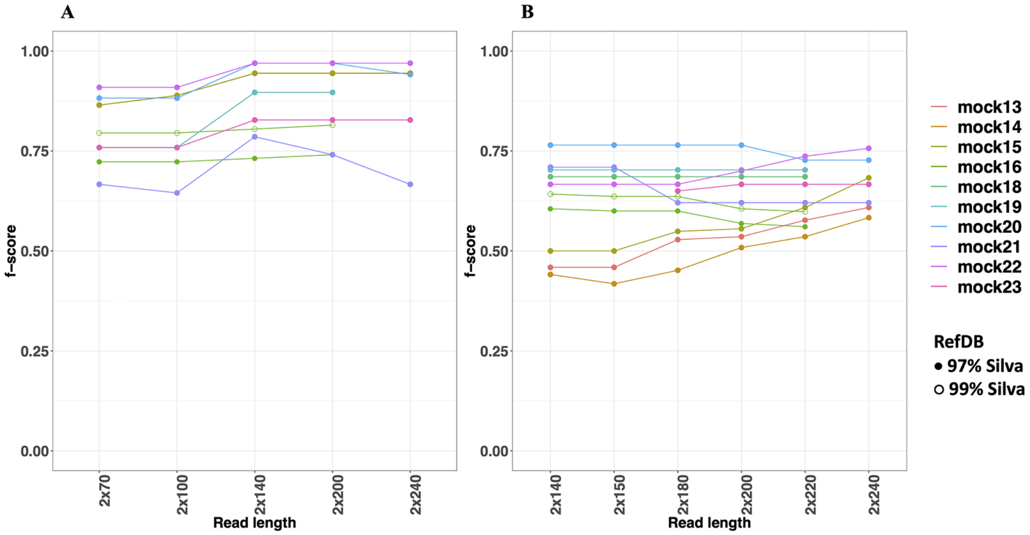


FIGURE 4 | F-scores of NG-Tax 2.0 **(A)** and DADA2 **(B)** at different read length. Silva 123 was used as reference database. The *x*-axis indicates the trimmed read length of the forward and reverse read. Note that mock16, mock 18 and mock 19 were not included in the comparison of the 240nt read length as after removal of primer sequences these reads were too short.

this respect may depend on the sample composition. We therefore selected a fixed read length per tool for further analysis: 140nt for NG-Tax 2.0 and 220nt for DADA2 as they provide the highest mean of the F-score calculated from all the communities at that length, which results in 0.89 and 0.64 respectively.

The two factors that contribute to the F-score are recall and precision. Both can be used to assess the quality of the pipeline and are equally important. In general, the level of recall of DADA2 and NG-Tax 2.0 were comparable with an average of 0.77 and 0.85, respectively. However, the precision of NG-Tax 2.0 was noticeably higher than that of DADA2 with an average of 0.95 vs 0.58 (**Figure 5**). The results show that both tools are equally good at inferring the expected microbial composition from the sample. However, DADA2 tended to predict taxonomic assignment of a higher rank, which led to a lower precision and F-score. Similar results using two staggered mocks from Tourlousse et al., 2017 with two replicates each can be found in **Data Sheet 3**.

Modified Rv Coefficient

An alternative metric used to determine the efficiency of both pipelines is the modified RV coefficient. Unlike the previous statistical measures, the modified RV coefficient takes into account the relative abundance of the identified bacteria, which is crucial for understanding a pipeline's performance. **Figure 5** shows that the modified RV coefficient from NG-Tax 2.0 on both the number of taxonomic lineages and their corresponding relative abundances are closer to the actual composition than DADA2. The average for NG-Tax 2.0 is 0.74 whereas the average coefficient for DADA2 is 0.28.

Tracking of Asvs Across Multiple Samples

ASVs have a single nucleotide resolution and are assumed to be directly derived from an existing biological entity. As in NG-Tax, ASV objects contain the forward and reverse sequence of the specific ASV (**Figure 2**), we can design SPARQL queries to explore the presence of specific ASVs across multiple mock samples. As most of the selected mocks are not biologically related, the majority of the ASVs will only be present in a single sample. Mock13-15, however, are composed of genomic DNA from the same 21 bacterial isolates and thus we expected a high number of ASVs shared between these three samples. The composition of mock13-15 includes three *Streptococcus* species being *Streptococcus agalactiae* ATCC BAA-611, *Streptococcus mutans* ATCC 700610, and *Streptococcus pneumoniae* ATCC BAA-334, each of which has multiple, but not necessarily identical copies of the 16S rRNA gene. For instance, the *Streptococcus agalactiae* genome contains seven copies of the 16S rRNA gene. Nine distinctive mock13 ASV objects are taxonomically annotated as *Streptococcus* and amplicon sequences linked to five of those objects showed 100% sequence identity with separate *Streptococcus agalactiae* 16S rRNA genes. A SPARQL query showed that four of these ASVs are present in all three mocks while one is not present in mock14. Overall, of the 60 taxonomically annotated ASVs in mock13, 56 variant sequences are present in all three mocks. Similarly, when we include in the query the unrelated mock16 composed of genomic DNA from 57 bacterial isolates, the expected taxon overlap is four; *Bacteroides*, *Porphyromonas*, *Deinococcus* and *Enterococcus*. The SPARQL query showed that five distinct ASVs are present in all four mocks. Two ASV's were annotated as *Bacteroides*, the other three as

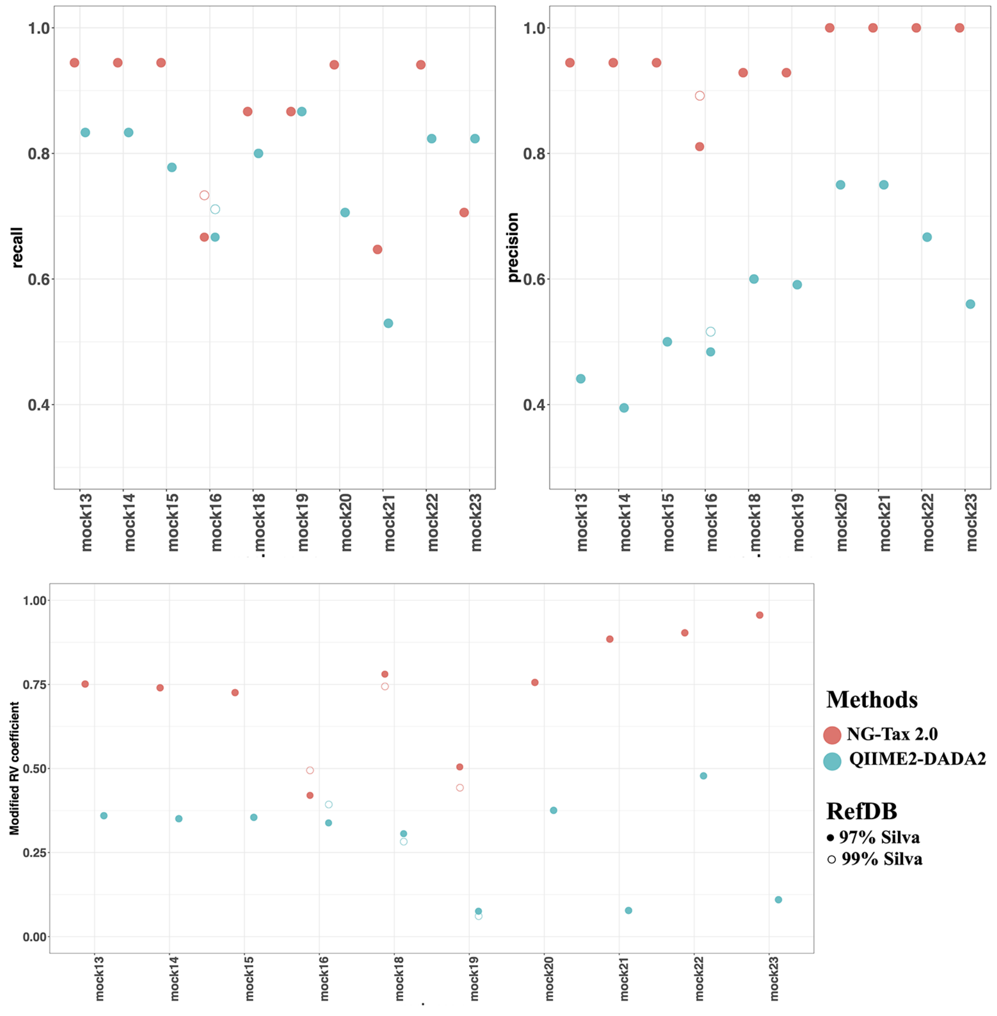


FIGURE 5 | Recall, precision and modified RV coefficient of NG-Tax 2.0 and DADA2. NG-Tax 2.0 is labelled in *red* and DADA2 is labelled in *blue*. *Upper panel* left, recall; right, precision. *Lower panel* modified RV coefficient. Silva 123 is used as reference database clustered at 97 (*filled circles*) and 99% (*open circles*). Note that in many cases results overlap in which case only the results obtained with the 97% threshold is shown.

Porphyromonas, Deinococcus and *Enterococcus*. **Figure 6** summarizes the result of a SPARQL query for the presence of specific ASVs amplicon sequence variants across all mocks.

Impact of Incremental Databases

The taxonomic annotation of a 16S rRNA gene amplicon depends on many variables, including the version of the reference database used. Because new phylogenetic groups are constantly being discovered (Hug et al., 2016), obtaining a correct bacterial phylogeny will remain a moving target for some time. Hence, keeping track of how the amplicon data was analysed, the data provenance, is critical. The observation that even a single reference database, clustered at two different similarity thresholds can lead to different results led us to investigate the impact of incremental versions of the SILVA database. For this, we used SPARQL queries to analyse the taxonomic annotation of the ten selected mock communities using three incremental stable versions of the SILVA database, namely releases 123, 128 and 132. NG-Tax 2.0 has the ability to create a custom taxonomic reference file *de novo* using a set of unaligned reference sequences as input. This allows for instance to add a new species to an existing taxonomic reference file. To demonstrate this feature we built a custom reference file using 16S rRNA gene sequences obtained from Hug et al., 2016. The SILVA result showed that in the latest version of the SILVA database some taxa have been reclassified. For instance, in mock18 the phylum and class of *Treponema_2* have been reclassified from *Spirochaetae* and *Spirochaetes* to *Spirochaetes and Spirochaetia*. The class and order of *Nitrosomonas* were also reclassified from Betaproteobacteria and *Nitrosomonadales* to Gammaproteobacteria and *Betaproteobacteriales*. Not unexpected the biggest “change” was when we compared taxonomic reference files from different origins. Results are summarized in **Table S1**.

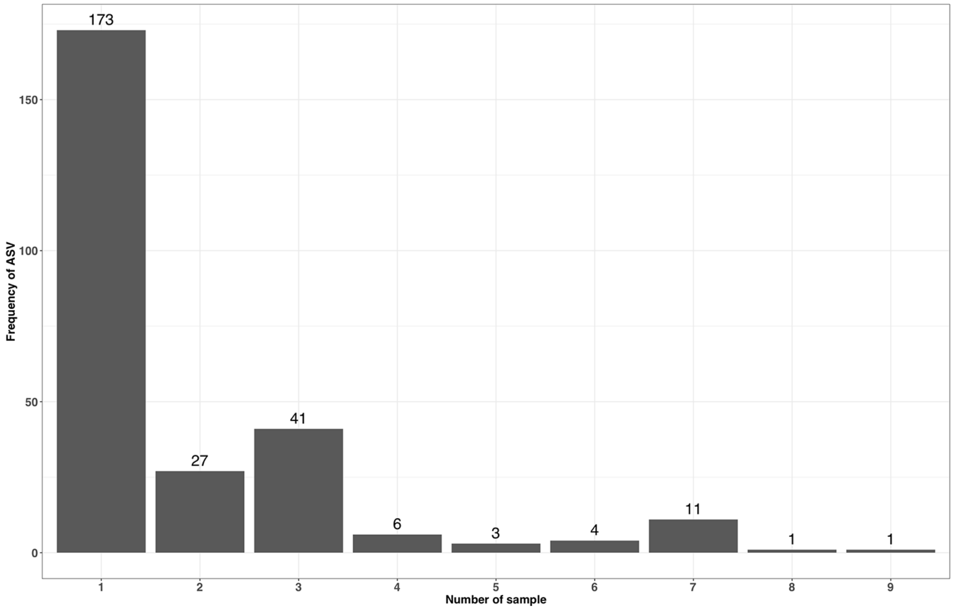


FIGURE 6 | Occurrence of accepted ASV forward and reverse sequences with a read length of 70 across multiple mock samples.

DISCUSSION

NG-Tax 2.0 is an open software framework that uses semantic technologies for data and knowledge management. It is particularly designed for FAIR and high-throughput taxonomic classification and downstream analysis of marker gene amplicon sequences. By using the RDF data model, NG-Tax 2.0 is able to engage a traceable *de novo* OTU picking and de-noising algorithm, generating ASV objects that link ASV sequence data with the full data-wise and element-wise provenance. The linked data structure ensures a high degree of interoperability. Serialized ASV objects can be automatically stored in a standard graph database structure and directly queried for comparative analyses of data and meta-data across thousands of samples.

For targeted amplicon sequencing, denoising, i.e. the separation of biological variation from amplicon sequencing errors, is essential to increase the reliability of downstream analyses. Clustering sequences into OTUs has been routinely applied in the past to reduce the impact of sequence errors and to speed up the analysis process by picking a representative sequence (Nguyen et al., 2016). However, many recent studies now use a 100% similarity threshold or ASVs. ASVs are standardly generated with NG-Tax 2.0 and with DADA2, one of the most commonly used pipelines today. As both NG-Tax 2.0 and DADA2 have a single nucleotide resolution, the number of ASVs and taxonomic annotation from NG-Tax 2.0 and DADA2 should be the same, however, the specific criteria used to remove erroneous-sequences creates the differences.

To test the performance of NG-Tax 2.0 we used ten 16S rRNA gene mock communities, staggered and even, and compared the results with those obtained with DADA2. We showed that while the recall of the expected microbial composition for both pipelines was comparable, there are substantial differences in the precision and the prediction of relative abundances. We proposed the use of a modified RV coefficient for evaluating the performance of a given pipeline (Smilde et al., 2009). It measures the common information of two matrixes which represent the relative abundance distributions of the microbial composition. This increases the efficiency in differentiating between two communities as compared to the binary classifier. The advantage in using this method is the ease of interpretation. The results are presented as a single value, which is convenient for visualization, and it can be interpreted in the same way as a correlation coefficient with the value between -1 and 1, which is already familiar to biologists.

Discussions about how to analyse microbial community data is an on-going process, and the golden standard for microbiome analysis has not yet been settled (Knight et al., 2018; Pollock et al., 2018). DADA2 generates a parametric error model based on the dataset and uses it to remove or collapse the sequences. On the other hand, NG-Tax 2.0 employs an empirically determined relative abundance cut-off taking into account the evenness of the read distribution over the ASVs to flag ASVs with an associated low read count that are considered as artefacts. It then attempts to merge those artefacts with ASVs with high read counts, which are more likely to be true ASVs, using a single mismatch as criterium. While both methods seem to be effective in recalling the expected composition, precision of NG-Tax 2.0 was much higher than that of DADA2 mainly because the parametric model predicted more ASVs, an effect that will increase along with the diversity of the community (Nearing et al., 2018).

NG-Tax 2.0's novelty is in using the RDF data model to transform amplicon data into ASV objects that link ASV sequences data with the dataset-wise and element-wise provenance. This not only greatly enhances the reproducibility of the analysis but also increases the degree of interoperability of the data required for comparative analyses. For instance, in finding rare species in a particular community, DADA2 may have the advantage while at the same time risking that those organisms are artefacts. In NG-Tax 2.0, rejected ASVs with relatively low read abundances are flagged as artefacts but due to a high degree of interoperability NG-Tax 2.0 enables a reanalysis of the data by comparing them between multiple samples and by using alternative parameter settings.

To conclude, NG-Tax 2.0 provides a simple to use, semantic framework for high-throughput microbiota analysis. Due to use of the RDF data model it allows to generate fully traceable ASV objects that link ASV sequence data with the full data-wise and element-wise provenance. This data model allows users to systematically adjust the parameters for the reanalysis or infer the biology behind these sequences using comparative analyses.

We compared the analysis results from the publicly available mock communities against those obtained by DADA2. The outcome shows that both pipelines are able to recall the microbial composition from the reference. However, NG-Tax 2.0 shows a higher precision score and the predicted relative abundances are closer to the expected composition than those provided by DADA2.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JD, WP, PS, and JK developed the code. WP, PS, and JK performed the computational analyses. WP and GH wrote the original draft of the manuscript. WP, GH, JD, JK, HS, and PS contributed to the writing, review, and editing of the manuscript.

FUNDING

WP is financially supported by a Royal Thai Government Scholarship, Thailand. This work has received funding from the Netherlands Organisation for Scientific Research funded UNLOCK project (NRGWI.obrug.2018.005) and from the European Union H2020 under grant agreement No. 730976 (IBISBA 1.0).

ACKNOWLEDGMENTS

We thank Koen Venema and Marisol Aguirre for use of the data presented in **Supplementary Material 3**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01366/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10. doi: 10.1093/nar/gkw343

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., et al. (2016). mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 1 (5). doi: 10.1128/mSystems.00062-16

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

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

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, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2017). shiny: web application framework for R.

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

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341, 1237439. doi: 10.1126/science.1237439

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

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* 34, 942. doi: 10.1038/nbt.3601

Hall, M., and Beiko, R. G. (2018). "16S rRNA Gene Analysis with QIIME2," in *Microbiome analysis: methods and protocols*. Eds. R. G. Beiko, W. Hsiao and J. Parkinson (New York, NY: Springer New York), 113–129. doi: 10.1007/978-1-4939-8728-3_8

Hawkins, J. A., Jones, S. K., Finkelstein, I. J., and Press, W. H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. *Proc. Natl. Acad. Sci.* 115, E6217. doi: 10.1073/pnas.1802640115

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi: 10.1038/nmicrobiol.2016.48

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

Konstantinidis, K. T., and Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* 187, 6258–6264. doi: 10.1128/JB.187.18.6258-6264.2005

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/AEM.01043-13

Musen, M. A.the Team P. (2015). The Protégé Project: a look back and a look forward. *AI Matters* 1, 4–12. doi: 10.1145/2757001.2757003

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error–correction methods. *PeerJ* 6, e5364. doi: 10.7287/peerj.preprints.26566v1

Nguyen, N. P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2. doi: 10.1038/npjbiofilms.2016.4

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Appl. Environ. Microbiol.* 84. doi: 10.1128/AEM.02627-17

Quast, C., Pruesse, E., Gerken, J., Peplies, J., Yarza, P., Yilmaz, 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

Ramiro-Garcia, J., Hermes, G. D. A., Giatsis, C., Sipkema, D., Zoetendal, E. G., Schaap, P. J., et al. (2016). NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes. *F1000Research* 5, 1791. doi: 10.12688/f1000research.9227.2

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43, e37–e37. doi: 10.1093/nar/gku1341

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinf.* 17, 125. doi: 10.1186/s12859-016-0976-y

Shin, S., and Park, J. (2016). Characterization of sequence-specific errors in various next-generation sequencing systems. *Mol. Biosyst.* 12, 914–922. doi: 10.1039/C5MB00750J

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and Van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics* 25, 401–405. doi: 10.1093/bioinformatics/btn634

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849. doi: 10.1099/00207713-44-4-846

Stulberg, E., Fravel, D., Proctor, L. M., Murray, D. M., LoTempio, J., Chrisey, L., et al. (2016). An assessment of US microbiome research. *Nat. Microbiol.* 1, 15015. doi: 10.1038/nmicrobiol.2015.15

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117

Tourlousse, D. M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N., and Sekiguchi, Y. (2017). Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* 45. doi: 10.1093/nar/gkw984

Tringe, S. G., and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* 6, 805–814. doi: 10.1038/nrg1709

van Dam, J. C. J., Koehorst, J. J. J., Vik, J. O., Schaap, P. J., and Suarez-Diez, M. (2019). The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation. *Sci. Data* 6, 254. doi: 10.1038/s41597-019-0263-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3. doi: 10.1038/sdata.2016.18

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K.-H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Micro.* 12, 635–645. doi: 10.1038/nrmicro3330

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi: 10.1093/nar/gkt1209

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

HYPOTHESIS AND THEORY
published: 21 April 2020
doi: 10.3389/fgene.2020.00310

A Primer for Microbiome Time-Series Analysis

Ashley R. Coenen[1*†], *Sarah K. Hu*[2*†], *Elaine Luo*[3*†], *Daniel Muratore*[4*†] and *Joshua S. Weitz*[1,5*]

[1] School of Physics, Georgia Institute of Technology, Atlanta, GA, United States, [2] Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA, United States, [3] Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Honolulu, HI, United States, [4] Interdisciplinary Graduate Program in Quantitative Biosciences, Georgia Institute of Technology, Atlanta, GA, United States, [5] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, United States

OPEN ACCESS

Edited by:
*Joao Carlos Setubal,
University of São Paulo, Brazil*

Reviewed by:
*Ricardo Maria Letelier,
Oregon State University, United States
Karoline Faust,
KU Leuven, Belgium*

***Correspondence:**
*Ashley R. Coenen
acoenen3@gatech.edu
Sarah K. Hu
sarah.hu@whoi.edu
Elaine Luo
elaine.luo@hawaii.edu
Daniel Muratore
dmuratore3@gatech.edu
Joshua S. Weitz
jsweitz@gatech.edu*

† These authors have contributed equally to this work

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

***Received:** 29 May 2019*
***Accepted:** 16 March 2020*
***Published:** 21 April 2020*

Citation:
Coenen AR, Hu SK, Luo E, Muratore D and Weitz JS (2020) A Primer for Microbiome Time-Series Analysis. Front. Genet. 11:310. doi: 10.3389/fgene.2020.00310

Time-series can provide critical insights into the structure and function of microbial communities. The analysis of temporal data warrants statistical considerations, distinct from comparative microbiome studies, to address ecological questions. This primer identifies unique challenges and approaches for analyzing microbiome time-series. In doing so, we focus on (1) identifying compositionally similar samples, (2) inferring putative interactions among populations, and (3) detecting periodic signals. We connect theory, code and data via a series of hands-on modules with a motivating biological question centered on marine microbial ecology. The topics of the modules include characterizing shifts in community structure and activity, identifying expression levels with a diel periodic signal, and identifying putative interactions within a complex community. Modules are presented as self-contained, open-access, interactive tutorials in R and Matlab. Throughout, we highlight statistical considerations for dealing with autocorrelated and compositional data, with an eye to improving the robustness of inferences from microbiome time-series. In doing so, we hope that this primer helps to broaden the use of time-series analytic methods within the microbial ecology research community.

Keywords: microbial ecology, time-series analysis, marine microbiology, inference, clustering, periodicity, code:R, code:matlab

1. INTRODUCTION

Microbiomes encompass biological complexity from molecules to genes, metabolisms, and community ecological interactions. Understanding this complexity can be difficult due to domain- or location- specific challenges in sampling and measurement. The application of sequencing technology has revolutionized almost all disciplines of microbial ecology, by allowing researchers the opportunity to access the diversity, functional capability, evolutionary history, and spatiotemporal dynamics of microbial communities rapidly and at a new level of detail (Huse et al., 2008; Caron, 2013). Increasingly it is now possible to sample at the time-scale at which those processes occur, resulting in the collection of microbiome time-series data. While such high-resolution sampling opens new avenues of inquiry, it also presents new challenges for analysis (McMurdie and Holmes, 2014; Weiss et al., 2016, 2017; Widder et al., 2016; Knight et al., 2018).

One of the first challenges in analyzing microbiome data is to categorize sequences in terms of taxa or even "species" (Konstantinidis et al., 2006; Caron and Hu, 2019). Many methods have been developed to perform this categorization (Blaxter et al., 2005; Konstantinidis and Tiedje, 2005; Huse et al., 2008; Mende et al., 2013; Sunagawa et al., 2013; Eren et al., 2014; Katsonis et al., 2014; Mahé et al., 2015; Varghese et al., 2015; Roux et al., 2016; Callahan et al., 2017; Luo et al., 2017). Particular choices used to define species-level units may alter downstream estimations of diversity and other parameters of interest (Youssef et al., 2009; Kim et al., 2011; Hu et al., 2015). Indeed, even the procedures for estimating common diversity parameters are impacted by the properties of read count data (Willis, 2019). However, some definition of taxa is often necessary for characterizing the composition of microbial communities. In this primer, we use the term *taxon* to denote approximately species-level designations, such as operational taxonomic unit (OTU) or amplicon sequence variant (ASV).

Once sequences have been categorized to approximate species-level groups, the interpretation of their read count abundances is accompanied by assumptions that violate many standard parametric statistical analyses. For example, zero reads from a sample mapping to a particular taxon is commonplace in microbiome sequence results, yet it typically remains unclear if a zero indicates evidence of absence (e.g., taxon not present in sample, incapable of transcribing a gene) or absence of evidence (e.g., below detection, inadequate sequencing depth) (Paulson et al., 2013; Weiss et al., 2017). In addition, sequence data is compositional, and therefore does not include information on absolute abundances (Gloor et al., 2017). As a result, compositional data has an intrinsic negative correlation structure, meaning that the increase in relative abundance of one community member necessarily decreases the relative abundances of all other members (Silverman et al., 2017).

The issues of categorization and sampling depth apply to all kinds of microbiome data sets. In particular, temporal autocorrelation presents an additional complexity to microbiome time-series, in that each observation is dependent on the observations previous to it in time. Autocorrelation also precludes the use of many standard statistical techniques, which assume that observations are independent. In **Figure 1**, we show how autocorrelation leads to high incidences of spurious correlations among independent time-series.

Complex microbiome data demand nuanced analysis. In this paper, we provide a condensed synthesis of principles to guide microbiome time-series analysis in practice. This synthesis builds upon and is complementary to prior efforts that established the importance of analyzing temporal variation for understanding microbial communities (e.g.,

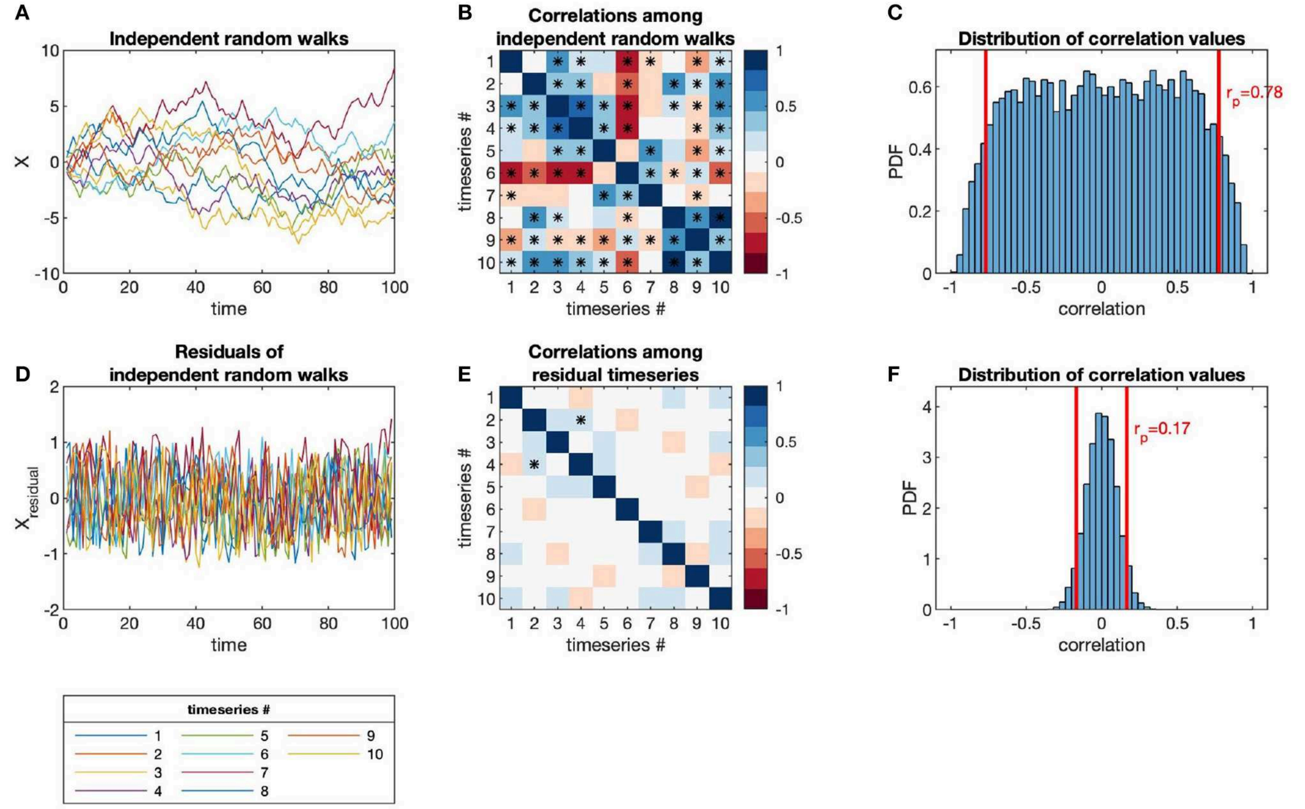


FIGURE 1 | Independent random walks yield apparently significant correlations (when evaluated as independent pairs) despite no underlying interactions, in contrast to residuals (i.e., point-to-point differences). **(A)** Time-series of independent random walks, $x_i(t)$. **(B)** Correlation structure of independent random walks. **(C)** Distribution of correlation values for an ensemble of independent random walks, with p-value = 0.05 marked (red lines). **(D)** Time-series of the residuals of independent random walks, i.e., $\Delta x_i(t) = x_i(t + \Delta t) - x_i(t)$. **(E)** Correlation structure of residual time-series. **(F)** Distribution of correlation values for the same ensemble as **(C)** but taken between the residual time-series, with p-value = 0.05 marked (red lines).

Faust et al., 2015). Here, we introduce core statistical methods for microbiome time-series analysis as a starting point and suggest further reading on other possible methods. Our process is described in detail via several code tutorials at https://github.com/WeitzGroup/analyzing_microbiome_timeseries that include analytic tools and microbiome time-series data, and provide a software skeleton for the custom analysis of microbiome time-series data. These tutorials include the basics of discovering underlying structure in high-dimensional data via statistical ordination and divisive clustering, non-parametric periodic signal detection in temporal data, and model-based inference of interaction networks using microbiome time-series.

2. METHODS

2.1. Overview of Tutorials

We describe three distinct categories of time-series analyses: clustering, identifying periodicity, and inferring interactions. For each category, we demonstrate analyses that answer an ecologically motivated question (**Figure 2**). Each tutorial emphasizes normalization methods specifically developed for the analysis of compositional data. Each tutorial also addresses challenges related to multiple hypothesis testing, overdetermination, and measurement noise. Interactive, self-contained tutorials that execute the workflows described in the manuscript are available in

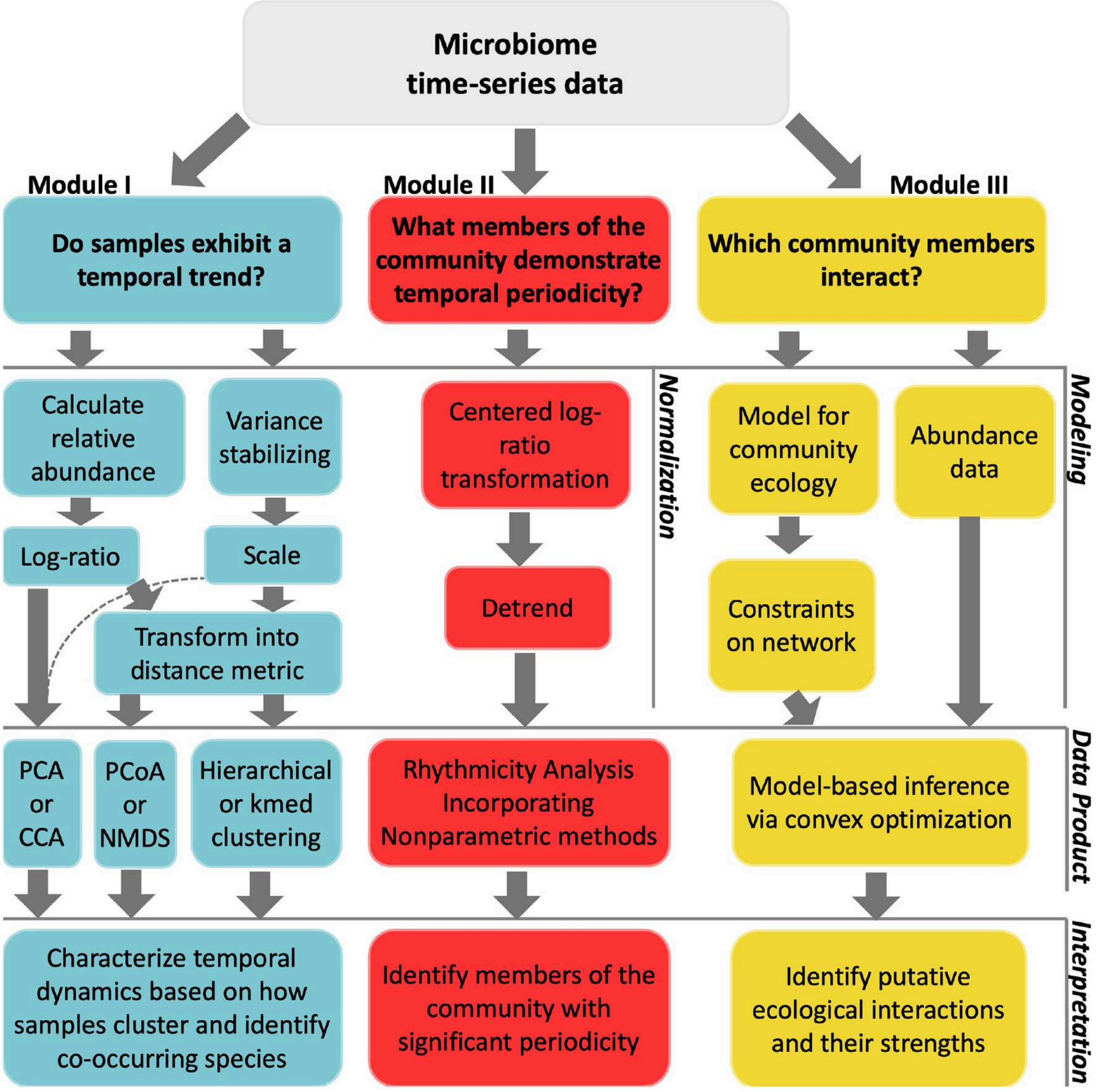

Microbiome time-series data
Module I
Module II
Module III
Do samples exhibit a temporal trend?
What members of the community demonstrate temporal periodicity?
Which community members interact?
Normalization
Modeling
Calculate relative abundance
Variance stabilizing
Log-ratio
Scale
Transform into distance metric
Centered log-ratio transformation
Detrend
Model for community ecology
Abundance data
Constraints on network
Data Product
PCA or CCA
PCoA or NMDS
Hierarchical or kmed clustering
Rhythmicity Analysis Incorporating Nonparametric methods
Model-based inference via convex optimization
Interpretation
Characterize temporal dynamics based on how samples cluster and identify co-occurring species
Identify members of the community with significant periodicity
Identify putative ecological interactions and their strengths

FIGURE 2 | Workflow of techniques implemented in each module. The top layer considers questions of interest for a particular study. In the second layer, data normalizations are listed as implemented in module I and module II. For module III, we use synthetic data and instead list modeling inputs. The third layer shows the analytical techniques used in this primer, which we note is not exhaustive. These techniques provide some insight into the initial question asked, as described in the fourth layer.

R and Matlab https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

2.2. Dataset Sources

For modules I and II, time-series data are derived from an 18S rRNA gene amplicon data set from Hu et al. (2018), in which samples were collected at 4 h intervals for a total of 19 time points (Lagrangian sampling approach). Input data are in the form of sequence count tables, where samples are represented as columns and each row is a taxonomic designation (OTU or transcript ID) with sequence counts or read coverage abundance per taxon (here we use "taxon" as shorthand). The code in each of these modules can be customized for use on other data, although for the purposes of analyzing any temporal-scale variability, samples must be taken at a frequency sufficiently shorter than the temporal scale of interest (e.g., daily temporal variability requires sub-daily sampling, seasonal temporal variability requires sub-seasonal sampling).

For module III, time-series data are simulated from a synthetic microbial community, for which the "true" network is known. The techniques in this module can be applied to time-series data as has been done in a handful of studies (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018).

2.3. Normalization

2.3.1. Log-Ratio Transformations

Microbiome data tend to have three properties: (1) they are sum-constrained (all reads sum to the sequencing depth), (2) they are non-negative, and (3) they are prone to heteroskedasticity (the variance of the data is not equal across its dynamic range). These attributes of microbiome data violate some underlying assumptions of traditional statistical techniques. Transforming microbiome data into log-ratios (Aitchison, 1983) can mitigate these problems by stabilizing variance and distributing values over all real numbers, as well as mitigating statistical bias related to sequencing protocols (McLaren et al., 2019).

The simplest log-ratio transformation requires selecting some particular focal variable/taxon in the composition, dividing all other variables in each sample by the abundance of the focal taxon, and taking the natural logarithm. Mathematically:

$$LR_i = ln(x_i) - ln(x_{focal}) \tag{1}$$

This kind of log-ratio transformation eliminates negative constrained covariances, but all variables become relative to the abundance of an arbitrary focal taxon. Instead of selecting a focal taxon, the *Centered Log-Ratio Transformation* constructs ratios against the geometric average of community abundances (Egozcue et al., 2003).

$$CLR_i = ln(x_i) - \frac{1}{n}\sum_{k=1}^{n} ln(x_k) \tag{2}$$

This transformation retains the same dimensionality as the original data, but is also still sum constrained:

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n}\left(ln(x_k) - \frac{1}{n}\sum_{k=1}^{n} ln(x_k)\right) \tag{3}$$

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n} ln(x_k) - \frac{n}{n}\sum_{k=1}^{n} ln(x_k) \tag{4}$$

$$= 0 \tag{5}$$

Log-based transformations require some caution when dealing with data sets with large numbers of zeros, namely because the logarithm of zero is undefined. To overcome this problem, implementations usually employ some pseudocount method, i.e., adding a small number to all observations to make the log of zero observations calculable. Adding a pseudocount disproportionately affects rare taxa, where the magnitudes of differences between samples may be similar to the magnitude of the added pseudocount and therefore obscured (Tsilimigras and Fodor, 2016).

2.3.2. Z-Score Transformation

Another transformation that converts data from counts to a continuous real-valued number is the z-score transformation, achieved by applying this relationship:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{6}$$

where x_i is an observation, μ_x is the mean of population x, and σ_x is the standard deviation of x. Often, μ_x and σ_x are estimated by the sample mean and standard deviation. The z-score is how far, in terms of number of standard deviations, a given observation is from the sample mean Cheadle et al. (2003). Of note, this transformation places variables of different magnitudes on a scale with the same range.

2.3.3. Variance Stabilizing Transformation

Log-ratio-based transformations in microbiome applications broadly serve the purpose of making the data more compatible with statistical methods that assume continuous/real-valued data and errors with equal variances. Such transformations are necessary because of the heteroscedasticity of sequence count data. A different approach to circumvent heteroscedastic data is to directly estimate a function which describes how the variance in the data increases as a function of the mean. Alternatively, it is possible to use a variance-stabilizing transformation, e.g., as implemented by the DESeq2 software package (Love et al., 2014). While the variance-stabilizing transformation is similar to a log transformation in the case of large counts, it is better suited to deal with zeros and does not rely on a pseudocount.

2.3.4. Distance Metric

Multivariate microbiome data is not necessarily easy to summarize or visualize in two or three dimensions. Therefore, to summarize and explore data, we want to recapitulate the high-dimensional properties of the data in fewer dimensions. Such low-dimensional representations are distance-based. A distance matrix is obtained by applying a distance metric to all pairwise combinations of observations. For example, given data matrix X, the Euclidean distance between observations X_i and X_j is:

$$d(X)_{ij} = \sqrt{(x_i - x_j)^2} \tag{7}$$

Different metrics measure distance using different attributes of the data [for comprehensive reviews of ecological distance metrics we recommend (Kuczynski et al., 2010; Buttigieg and Ramette, 2014)]. For example, only presence/absence of different community members is used to calculate Jaccard distance (Jaccard, 1912) and unweighted Unifrac (Lozupone and Knight, 2005), which also takes into account phylogenetic relationships between taxa. These metrics can be calculated on count data without transformation, and capture changes in the presence of rare taxa. On the other hand, Euclidean distance emphasizes changes in relative composition. Weighted Unifrac distance incorporates phylogenetic information as well as changes in relative abundances. Euclidean distance performed on log-ratio transformed data is analogous to Aitchinson's distance (Aitchison et al., 2000), which is recommended for the analysis of the difference of compositions.

In addition to distance metrics, sample-to-sample difference can also be compared by dissimilarities, such as the Bray-Curtis dissimilarity, which is defined between sample i and sample j as:

$$BC_{ij} = 1 - \frac{2\sum_{k=1}^{n} \min(s_{i,k}, s_{j,k})}{\sum_{k=1}^{n} s_{i,k} + \sum_{k=1}^{n} s_{j,k}} \tag{8}$$

where n is the total number of unique taxon observed between both samples, and $s_{i,k}$ is the abundance of taxon k in sample i. Bray-Curtis is widely used in ecological studies to measure differences in community composition (Bray and Curtis, 1957). A dissimilarity score of 0 means the two samples had identical communities, and a dissimilarity score of 1 means the two samples had no taxa in common. However, Bray-Curtis dissimilarity does not obey the triangle inequality (Gower and Legendre, 1986), which means that multivariate methods that assume distance matrices as input (e.g., NMDS) may yield uninterpretable results. For example, two samples that each have a Bray-Curtis dissimilarity of 0.05 from a third sample may have a Bray-Curtis dissimilarity of 1 from each other.

2.4. Ordination

2.4.1. Covariance-Based Ordination

Statistical ordination can be used to explore multivariate microbiome data. An ordination is a transformation that presents data in a new coordinate system, e.g., making high-dimensional data visualizable in two or three dimensions. Principal Components Analysis (PCA) is a method which selects this coordinate system via the eigen decomposition of the sample covariance matrix, i.e., which is equivalent to solving the factorization problem:

$$Q_{m\times m} = U_{m\times m} D_{m\times m} U_{m\times m}^T. \tag{9}$$

Here, Q is the sample by sample covariance matrix, D is a diagonal matrix containing the eigenvalues of Q, and U is a matrix of the eigenvectors associated with those eigenvalues. For PCA, the eigenvectors (or principal axes) are interpreted as new, uncorrelated variables, which are an orthogonal linear combination of the original m variables (Hotelling, 1933). Each of the eigenvalues corresponds to one of the eigenvectors and refers to its magnitude, which is proportional to the amount of variance in the data explained by that eigenvector. To plot a PCA, we select a subset of eigenvectors with the largest associated eigenvalues, apply the linear combination of variables contained in those eigenvectors to each observation, and then plot the observations with the resulting coordinates. Importantly, basic PCA relies on a least-squares approach for solving a linear model with the observed variables, which poorly models heteroscedastic non-negative data, such as taxon sequence counts. Non-linear PCA (Kramer, 1991) is one extension of PCA that can discover more sophisticated correlation structure between observed variables.

Principal Coordinates Analysis (PCoA), based on PCA, is another technique that allows for more flexibility in ordination modeling (Buttigieg and Ramette, 2014; Gloor et al., 2017). PCoA, on the other hand, uses the same procedure as PCA, except on a sample by sample distance matrix is decomposed instead of the sample covariance matrix (Borcard and Legendre, 2002), using the statistical properties of the distances instead of the original observed data. The choice of distance metric allows for the implementation of PCoA on either transformed (in which distance, such as euclidean may be suitable) or raw count (in which distance, such as Jaccard or unweighted Unifrac may be suitable) microbiome data. For both PCA and PCoA, scaling the data, for example with a z-score transformation, is recommended so that no one variable disproportionately influences the ordination (Holmes and Huber, 2019).

2.4.2. Non-metric Multidimensional Scaling

Non-metric Multidimensional Scaling (NMDS) is an alternative ordination method which forces data to be projected into a pre-specified number of dimensions (Kruskal, 1964). NMDS projects high-dimensional data into a lower-dimensional space such that all pairwise distances between points are preserved. To implement NMDS, we solve the optimization problem:

$$\hat{X'} = \arg\min \|d(X) - d(X')\|_2 \tag{10}$$

where X is the original data matrix and X' is the data in the lower-dimensional space. Here d is a distance metric (see Distance section). Because the sum of pairwise distances is the quantity being minimized by NMDS, this method is strongly affected by outliers, so data should be examined for outliers prior to NMDS ordination. Additionally, unlike PCA and PCoA, where the new sample coordinates are directly related to the measured variables, NMDS coordinates have no meaning outside of their pairwise distances. Another important difference between NMDS and PCA is that the NMDS is enforced to fit the ordination to a fixed number of dimensions, which means the projection is not guaranteed to be a good fit. *Stress* (Kruskal, 1964) is the quantification of how well the NMDS projection recapitulates the distance structure of the original data:

$$Stress = \sqrt{\frac{\sum(d(X) - d(X'))^2}{\sum d(X)^2}} \tag{11}$$

The closer the stress is to 0, the better the NMDS performed.

2.4.3. Clustering

Clustering defines relationships between individual data points, identifying a collection of points that are more similar to each other than members of other groups. Many clustering algorithms have been developed for the analysis of time series data (comprehensively reviewed in Liao, 2005). These algorithms include hierarchical methods, such as agglomerative clustering and k-medoids (McMurdie and Holmes, 2014; Gülagiz and Sahin, 2017), topological methods, such as self-organizing maps (Kohonen, 1990; Kavanaugh et al., 2014),and density-based methods, such as the DBSCAN algorithm (Khan et al., 2014). As a working example, we implement two types of hierarchical distance-based clustering algorithms, the partitioning about medoids (PAM or k-medoid) algorithm (Kaufman and Rousseeuw, 2009), and hierarchical agglomerative clustering (Murtagh, 1985). A hierarchical clustering method is one which works by partitioning the data into groups with increasingly similar features. The number of groups to divide the taxa into is determined prior to calculation, which begs the question: how many groups? This question can be quantitatively assessed using several indices. A clustering algorithm can be implemented using a range of possible numbers of clusters, and then comparison of these indices will indicate which number has a high degree of fit without over-fitting. These indices can also be used to help choose between clustering algorithms.

One such index is sum of squared differences, which is related to the total amount of uniformity in all clusters, defined as LaTeX error this align should read:

$$SSE = \sum_{k=0}^{n_{\text{clusters}}} \sum_{i=0}^{n_{\text{members}}} \left(\overbrace{x_{i,k}}^{\text{Cluster member}} - \overbrace{c_k}^{\text{Cluster center}} \right)^2 \tag{12}$$

A common heuristic to identifying an optimal number of clusters is to plot SSE vs. k and look for where the curve "elbows," or where the decrease slows down (Liu et al., 2010; Gülagiz and Sahin, 2017) (see clustering tutorial).

Another way to evaluate the efficacy of clustering is via the Calinski-Harabasz index (Calinski and Harabasz, 1974), which is the ratio of the between-cluster squared distances to the within-cluster squared differences (Liu et al., 2010):

$$CH = \frac{\frac{B(x)}{k-1}}{\frac{W(x)}{n-k}} \tag{13}$$

where $B(x)$ is the between cluster sum of square differences, $W(x)$ is the within cluster sum of square differences, n is the number of taxa, and k is the number of clusters. This index accounts for the number of clusters the data are partitioned into as well as the overall variation in the data as a whole. A large value of CH indicates that the between-cluster differences are much higher than the average differences between the dynamics of any pair of taxa in the data, so a maximum value of CH indicates maximum clustering coherence.

The "Silhouette width" is another index which allows for fine-scale examination of the coherence of individual taxon to their cluster. Silhouette width is therefore helpful for identifying outliers in clusters (Liu et al., 2010). The silhouette width for any given clustering of data is calculated for each taxon by taking the ratio of the difference between that taxon's furthest in-cluster neighbor and nearest out-of-cluster neighbor to the maximum of the two, such that

$$SW_i = \frac{\overbrace{min(d(x_i, x_{j\notin C}))}^{\text{sum square diff out of cluster}} - \overbrace{max(d(x_i, x_{j\in C}))}^{\text{sum square diff in cluster}}}{max(min(d(x_i, x_{j\notin C})), max(d(x_i, x_{j\in C})))} \tag{14}$$

where C is all taxa in the cluster, and d is the sum square difference operator. The widths can range from -1 to 1. Silhouette widths above 0 indicate taxa which are closer to any of their in-cluster neighbors than any out-of-cluster taxa, so having as many taxa with silhouette widths above 0 as possible is desirable. Any taxon with particularly low silhouette widths compared to the rest of their in-cluster neighbors should be investigated as potential outliers.

2.5. Periodicity Analysis

Periodicity analysis reveals whether or not a signal exhibits a cyclical periodic change in abundance. Approaches to identifying periodic signals include parametric methods and non-parametric methods. The multi-taper method is an example of a parametric method, which uses autoregression to find periodic signals in low signal-to-noise data (Mann and Lees, 1996) (for a software implementation in R https://cran.r-project.org/web/packages/ssa/index.html). Other examples of parametric methods include harmonic regression (Yang and Su, 2010; Ottesen et al., 2014), methods based on frequency spectral decomposition (Yang et al., 2011), and a widely used (Aylward et al., 2017; Hughes et al., 2017; Wilson et al., 2017; Hu et al., 2018) non-parametric method, "Rhythmicity Analysis Incorporating Non-parametric methods" (RAIN) (Thaben and Westermark, 2014).

The RAIN method identifies significant periodic signals given a pre-specified period and sampling frequency. RAIN then conducts a series of Mann-Whitney U tests [rank-based difference of means (Mann and Whitney, 1947)] between time-points in the time-series over the course of one period. For example, one such series of tests might answer the question: are samples at hours 0, 24, 48 higher in rank than the samples at hours 4, 28 (Hotelling, 1933). Then, the sequence of ranks is examined to determine if there is a consistent rise and fall about a peak time. For this procedure to work, RAIN relies on the assumption that time-series are stationary, or have the same mean across all sampled periods. One way to normalize microbiome time-series to better fit this assumption is detrending, or regression normalization, which removes longer-term temporal effects, such as seasonality. A first approximation of non-stationary linear processes can be made by taking the linear regression of all time-points with time as the independent variable, then subtracting this regression from the time-series. This operation stabilizes the data to have a similar mean across all local windows.

In order to assess periodicity for an entire microbial community, we may conduct many hypothesis tests. The more tests that are performed at once, the higher the probability of finding a low p-value due to chance alone (Streiner, 2015). Some form of multiple testing correction is therefore encouraged. False Discovery Rate (FDR) based methods are recommended for high-throughput biological data over more stringent Familywise Error Rate corrections (Noble, 2009; Glickman et al., 2014). The method employed here is the Benjamini-Hochberg step-up procedure (Benjamini and Yekutieli, 2001) (for graphical demonstration see the "periodicity" tutorial in the associated software package). P-values are ranked from smallest to largest, and all null hypotheses are sequentially rejected until test k where:

$$p_k \geq \frac{k}{m}\alpha \tag{15}$$

where m is the total number of tests conducted, and α is the desired false discovery rate amongst rejected null hypotheses. Alternative p-value adjustment methods designed for sequencing data have been proposed (Conneely and Boehnke, 2007) which take into account correlation between tests, although simulations (Stevens et al., 2017) demonstrate that for moderate effect sizes, methods, such as Benjamini-Hochberg generally control false discoveries as expected, if not slightly more conservatively.

2.6. Inferring Interactions

2.6.1. Model Specification of Ecological Dynamics

Inferring interactions using a model-based approach requires the specification of ecological (or eco-evolutionary) dynamics. Model specification requires extensive knowledge of the system of interest. Furthermore, models can be specified at diffierent levels of abstraction regarding taxonomic resolution (e.g. Storch and Šizling, 2008) and biological mechanisms (e.g. Vincenzi et al., 2016), leading to challenges in interpretability (Cao et al., 2017). Alternatively, data-driven identification of dynamical systems is an active area of research (e.g. Brunton et al., 2016; Mangan et al., 2016, 2017), providinga possible way forward when an appropriate model is not known *a priori*.

Currently, widely used models include some variation of Lotka-Volterra dynamics where each taxon is represented as a population whose abundances vary in time given density-dependent feedback with other populations (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Here, we focus on a variant of this class of problem, i.e., virus-microbe dynamics.

The microbe-virus ecological dynamics are modeled via a system of differential equations

$$\dot{H_i} = r_i H_i \left(1 - \frac{1}{K}\sum_{i'}^{N_H} H_{i'}\right) - H_i \sum_{j}^{N_V} M_{ij}\phi_{ij}V_j \tag{16}$$

$$\dot{V_j} = V_j \sum_{i}^{N_H} M_{ij}\phi_{ij}\beta_{ij}H_i - m_j V_j \tag{17}$$

where H_i and V_j denote the densities of host (i.e., microbe) type i and virus type j as they change over time. There are N_H host types and N_V virus types, each defined by their life history traits: growth rate r_i for host type i, decay rate m_j for virus type j, and a community-wide host carrying capacity K. The interactions between hosts and viruses are modeled as antagonistic infections culminating in the lysis (i.e., death) of the host cell and release of new viruses. For each pair host type i and virus type j, the infection is quantified by the interaction coefficient M_{ij}, adsorption rate ϕ_{ij} and burst size β_{ij}. The interaction coefficient is either 1 (the virus infects the host) or 0 (the virus does not infect the host) (Jover et al., 2013; Korytowski and Smith, 2017).

We randomly sample the life history traits and interaction parameters such that they are biologically plausible and guarantee local coexistence of all host and virus types (as described in Jover et al., 2016). We simulate the time-series of the resulting dynamical system using ODE45 in Matlab.

2.6.2. Objective Function for Model-Based Inference

We seek the interaction network that minimizes the difference between observed dynamics in densities and those predicted by the dynamical model. We use the virus equations (Equation 17) to derive the objective function

$$\min \quad \left\| W - \left(\tilde{M}^T \; -\vec{m}\right)\begin{pmatrix} H \\ \vec{1} \end{pmatrix} \right\|_2 + \lambda \, ||\!\sim M||_1 \tag{18}$$

$$\text{subject to} \quad \tilde{M}_{ij} > 0 \tag{19}$$

$$m_j > 0 \tag{20}$$

where W_{jk} is the per-capita derivative estimate of virus type j at sampled time t_k, H_{ik} is the density of host type i at sampled time t_k, $\tilde{M}^T_{ij} = M_{ij}\phi_{ij}\beta_{ij}$ is the weighted infection coefficient between virus type j and host type i and m_j is the decay rate of virus type j (as described in Jover et al., 2016). We seek to estimate the unknown weighted infection network $\tilde{M}$, using sampled densities of hosts H and viruses W over time.

To prevent over-fitting, we introduce a hyper-parameter λ, which can be tuned to control the sparsity of the inferred network M. Other approaches can also be used to identifya balance between goodness of fit and model complexity, such as k-crossfold validation or information criterion (e.g. AIC). For an exampleof using k-crossfold validation, see Stein et al. (2013).

2.6.3. Interaction Inference via Convex Optimization

In practice, we can solve the minimization problem (Equation 20) and infer the interaction network $\tilde{M}$ using convex optimization. Convex optimization is a well-developed technology for efficiently and accurately solving minimization problems of a particular form which are guaranteed to have a global minimum. Here, we use a freely available third-party software package for Matlab available for download at http://cvxr.com/cvx/ (Grant and Boyd, 2008, 2014) (also available for implementation in Python at https://www.cvxpy.org Diamond and Boyd, 2016; Agrawal et al., 2018). The details of implementation are described in Jover et al. (2016) and in the accompanying code tutorial.

In addition to convex optimization, there are many methods for inferring the interaction network, and dynamical systems parameters in general, from time-series. Two recent examples include MCMC fitting (Thamatrakoln et al., 2019; Zobitz et al., 2011) and Tikhonov regularization Stein et al. (2013).

3. RESULTS AND DISCUSSION

3.1. Exploring Shifts in Daily Protistan Community Activity

The North Pacific Subtropical Gyre (NPSG) is widely studied as a model ocean ecosystem. Near the surface, the NPSG undergoes strong daily changes in light input. Abundant microorganisms in the NPSG surface community, such as the cyanobacteria *Prochlorococcus* and *Crocosphaera*, adapt metabolic activities, such as cell growth and division to particular times of day (Aylward et al., 2015; Ribalet et al., 2015; Wilson et al., 2017). However, the extent to which these daily cycles and the timings of particular metabolic activities extend to protistan members of the NPSG surface ecosystem remains less characterized. To this end, we examined an 18S rRNA gene diel dataset from a summer 2015 cruise sampled every 4 h for 3 days on a Lagrangian track near Station ALOHA (Hu et al., 2018). In this expedition, both rRNA and rDNA were sampled to explore differences in metabolic activity for particular community members at different times of day (Hu et al., 2016). Previous work (Hu et al., 2018) found shifts in the metabolically active protistan community, including phototrophic chlorophytes and haptophytes as well as parasitic Syndiniales.

In this analysis, we asked whether or not the metabolically active component of the microbial community is unique to different times of day. Therefore, we focused specifically on the 18S rRNA gene data as a proxy for overall functional activity of protistan taxa (Charvet et al., 2014; Hu et al., 2016; Xu et al., 2017). We used statistical ordination to explore underlying sample covariance. Samples that appear near each other in a statistical ordination have similar multivariate structure. In the clustering tutorial we present several methods for performing ordination, e.g., NMDS and PCoA (see Methods: Ordination). In **Figures 3B,C**, we construct a PCoA using Jaccard distance to emphasize changes in presence/absence of rRNA signatures, and find that the first 3 Principal Coordinates explain 64.76% of the variation amongst all samples. Samples from 2 PM and 6 AM strongly differentiate along the first coordinate axis, while samples at 10 AM settle between them. The ordination suggests that the taxa which are transcribing the 18S rRNA gene at 2 PM are fairly distinct from those transcribing at 6 AM, while 10 AM is intermediate between the two. We also constructed a corresponding NMDS ordination using the same distance matrix that we constrained to two dimensions. The pattern of separation between 2 and 6 PM is maintained in this projection, reinforcing its importance as an underlying structural feature of these data. Next, we constructed an additional PCoA ordination on the Euclidean distance matrix of isometric log-ratio transformed 18S rRNA counts (see clustering tutorial for implementation). We select the isometric log-ratio transformation to alleviate the constraint of compositionality and to scale the data to a similar range of magnitudes, making Euclidean distance a suitable choice of distance metric. As seen in the scree plot in **Figure 3E**, while the first Principal Coordinate explained about 25% of the variation between samples, the following four Principal Coordinates each explained around 5% of the variation. Despite the low proportion of total variance explained, strong separation emerges between 2 PM and 6 AM samples along the largest coordinate axis. This ordination qualitatively agrees with a corresponding NMDS ordination (**Figure 3D**) forced into two dimensions.

Noting the differences in active community members between 2 PM and 6 AM, we identified co-occurring taxa by clustering their temporal dynamics after variance-stabilization and scaling normalizations (see clustering tutorial for discussion). Based on comparisons of sum squared errors and the CH index introduced in Methods, we opted to divide the OTUs into eight clusters (**Figure 4** for composition and representative temporal signature, tutorial for details on cluster selection). After comparing cluster evaluation metrics for hierarchical agglomerative clustering and a k-medoids algorithm, we conducted this clustering with k-medoids (see clustering tutorial for implementation). This method allows us to identify the time-series of the median taxon for each cluster as a representative shape for the cluster's temporal dynamics. We observe 2 PM peaks associated with clusters 2, 3, 6, and 8 and increased nighttime expression levels in cluster 1. These temporal patterns coincide with those surmised during our exploratory ordination of the community sampled at each time point (where 2 PM and 6 AM samples formed distinct clusters, **Figure 3**). Upon closer inspection of cluster membership (bar plots in **Figure 4A**), we find cluster 3 contains 65/105 (62%) of haptophyte OTUs and 18/33 (55%) of archaeplastids, including members of chlorophyta.

These results suggest temporal niche partitioning within the complex protistan community, consistent with the findings of Hu et al. (2018). By clustering results with respect to temporal patterns, we were able to parse the complex community to reveal the identities of key taxonomic groups driving the observed temporal patterns. The taxonomic composition of cluster 3 was made up of haptophytes and chlorophytes. Photosynthetic chlorophytes have previously been found to be correlated with the light cycle (Poretsky et al., 2009; Aylward et al., 2015) and the temporal pattern found in Hu et al. (2018) was similar to the standardized expression level (**Figure 4B**), as was the inferred relative metabolic activity of haptophytes.

3.2. Identifying Protists With Diel Periodicity in 18S Expression Levels

The metabolic activity of microbes is a critical aspect of the basis of marine food webs (Karl, 2002). In the euphotic zone, microbial populations are inherently linked to the light cycle as the energy source for metabolism. Identifying diel patterns in protists is particularly interesting due to widespread mixotrophy, where a mixotroph may ingest prey during periods of limiting inorganic nutrients or light (Nygaard and Tobiesen, 1993; Finkel et al., 2009; McKie-Krisberg et al., 2015). Additionally, protistan species encompass a wide range of cell sizes, thus the synchronization of light among photoautotrophs may reflect species-specific differences in nutrient uptake strategies (Hein et al., 1995; Gerea et al., 2019). Based on the observation of sample differentiation between the middle of the day (2 PM) and dawn (6 AM) from exploratory ordination and clustering analyses described in 4.1, we further investigated the hypothesis that some protists may exhibit a 24-h periodicity in their 18S rRNA gene expression levels.

The high-resolution nature of the sequencing effort in this study enabled us to ask which members of the protistan community had 24-h periodic signals. Following normalization (CLR, Equation 2) and detrending to center mean expression levels across the entire time series (see Periodicity tutorial and Methods: Periodicity Analysis), we used RAIN to assess the periodic nature of each OTU over time. Results from RAIN analysis reported p-values for each OTU at the specified period as well as estimates of peak phase and shape. The null hypothesis tested by RAIN is that the observations do not consistently increase, then decrease (or vice-versa) once over the course of a period. Rejecting the null hypothesis, then, asserts a time-series has one peak during the specified period. To determine which OTUs were found to have significant periodicity we rejected the null hypothesis at 5% FDR level (Equation 13). **Figure 5** illustrates examples of two protistan OTUs with significant diel periodicity, a haptophyte and stramenopile. Trends in CLR normalized values for each OTU indicated that there was a repeated and temporally coordinated relative increased in the metabolic activity of both taxa at 2 PM (**Figure 5**). Both groups have previously been found to respond to day-night environmental cues, which is also supported by Hu et al. (2018).

Identities of OTUs found to have significant diel periodicity included taxa with known phototrophic and/or heterotrophic feeding strategies. This suggests that taxa with diel changes in metabolic activity may be responding to light or availability of prey. More specifically, several known phototrophs or mixotrophs, including dinoflagellates, haptophytes, and

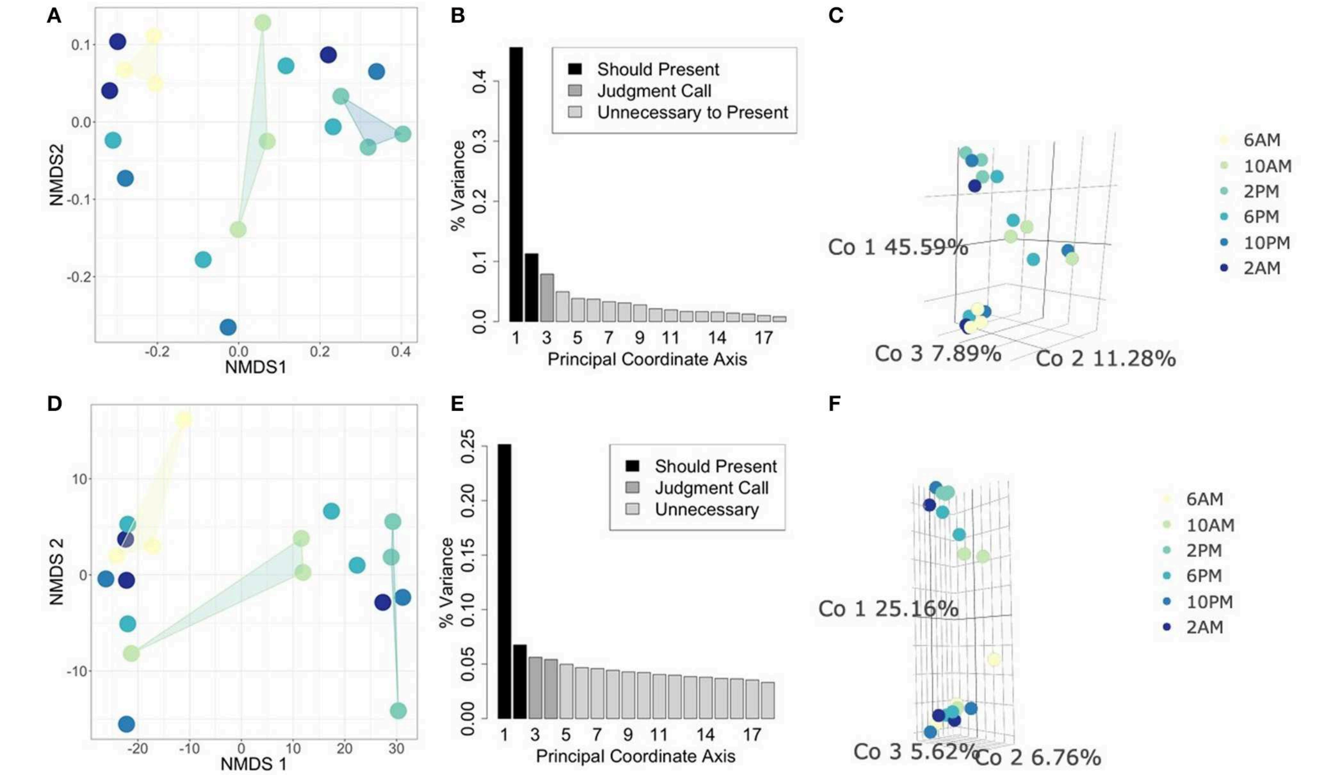


FIGURE 3 | Comparing statistical ordination techniques for 18S community compositions across samples. (Top row) Ordinations using Jaccard distance for comparison of presence/absence of community members between samples. (Bottom row) Ordinations using Euclidean distance on isometric log-ratio transformed data. **(A,D)** Non-metric Multidimensional Scaling (NMDS) projection in two dimensions, arbitrary units. Convex hulls have been drawn to emphasize ordinal separation of 6 AM (yellow), 10 AM (light green), and 2 PM (teal) samples. **(B,E)** Scree plots for PCoA ordinations. Each bar corresponds to one axis of the PCoA, the height is proportional to the amount of variance explained by that axis. We decided the first 3 axes were necessary to summarize the data in these cases [explaining a total of **(B)** 64.76% and **(E)** 37.54% of the variance]. Shading of bars indicate our interpretations of which axes are important to show (black), which are unimportant (light gray), and which are intermediate cases (medium gray). **(C,F)** PCoA ordinations using the selected axes after scree plot examination. Each point is one sample, the color of the point indicates the time of day at which the sample was taken (colors correspond to NMDS projections).

stramenopiles were found to have significant diel periodicity. Interestingly, there were a number of OTUs identified as belonging to the Syndiniales group (Alveolates) which are obligate parasites. Diel rhythmicity among these parasites suggests that they may be temporally coordinated to hosts that also have a periodic signal, which includes dinoflagellates.

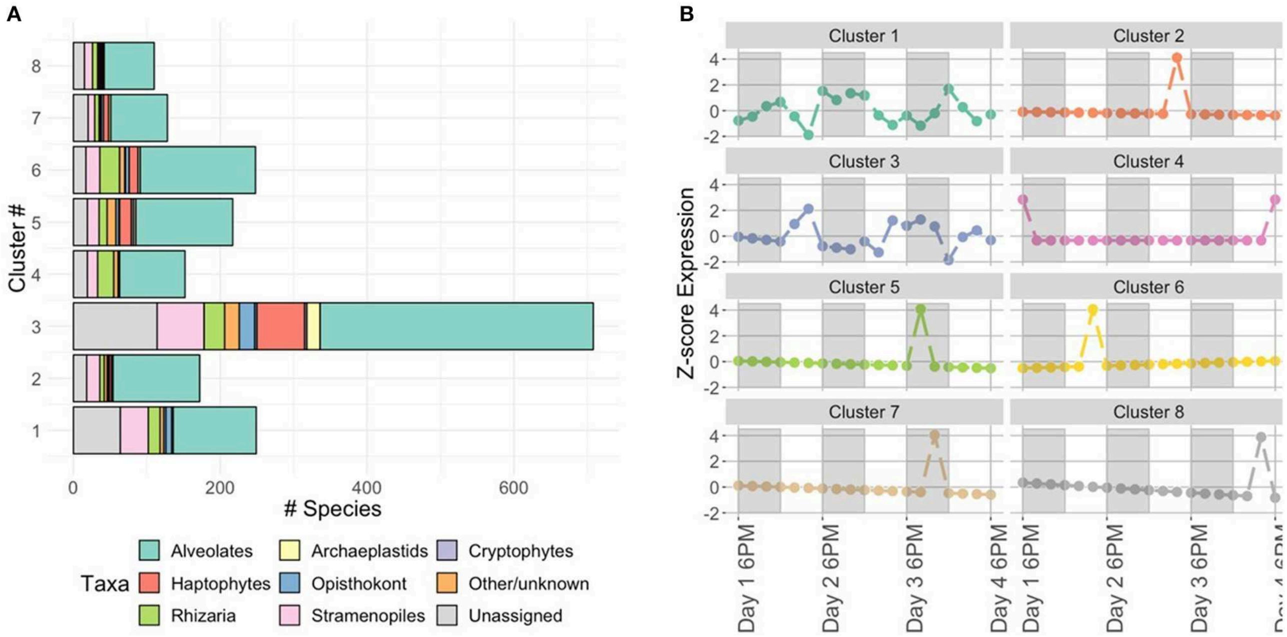


FIGURE 4 | Characterization of protist clusters. **(A)** Cluster membership based on the phylum or class level protistan taxonomy. The "Other/unknown" category includes sequences with non-specific identity, such as "uncultured eukaryote" and "Unassigned" denotes sequences with no taxonomic hit (< 90% similar to reference database). **(B)** Representative taxon time-series for each cluster. Y-axis is z-score (see Methods: Normalizations), so a value of 0 corresponds to mean expression level. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).

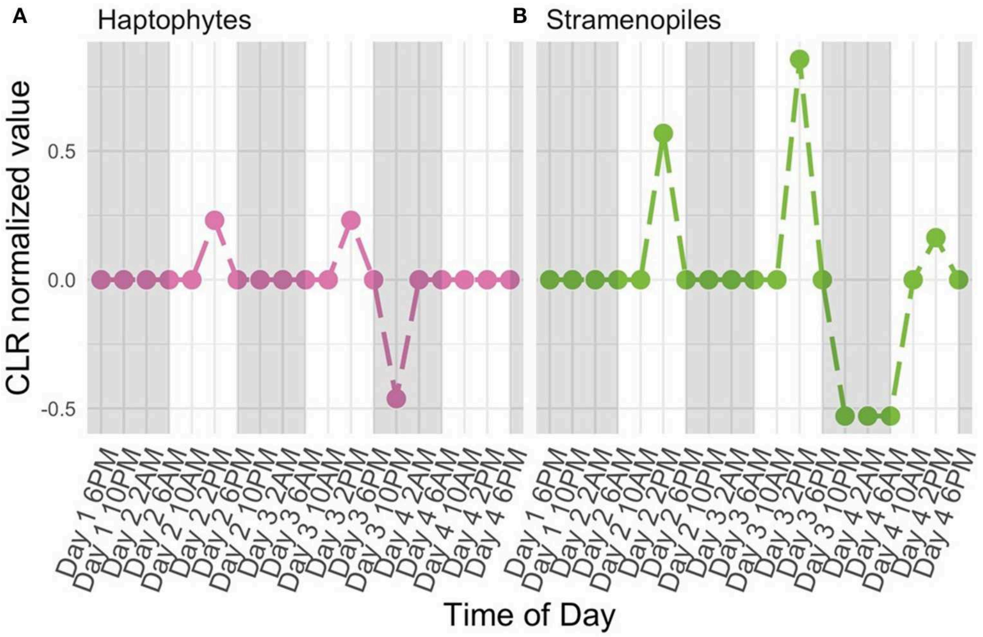


FIGURE 5 | Centered Log Ratio (CLR)-transformed, detrended 18S rRNA gene levels (y-axes) over time (x-axes) for a subset of OTUs found to have significant diel periodicity (RAIN analysis). A value of 0 denotes the mean expression level for a given OTU. Included OTUs belong to the **(A)** Haptophyte and **(B)** Stramenopile groups. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).

3.3. Inferring Interactions in a Synthetic Microbial Community

The goal of an inference method is to quantify ecological interactions between pairs of populations or taxonomic designation of interest. The result of such analysis is an interaction network for the community of interest. In the context of microbial communities, "interaction" can be broadly defined and include, for example, direct competition for a nutrient, toxin-mediated attacks, or cooperation via exchange of secondary metabolites. Besides pairwise interactions between microbes, other interactions may be of interest, such as higher-order interactions [e.g., three-way microbial exchanges (Fisher and Mehta, 2014; Bairey et al., 2016; Grilli et al., 2017)], pressures from other trophic levels (e.g., grazers, viruses), or driving via environmental variables (e.g., antibiotics, nutrient flux). Inferring interaction networks is a challenging task, in part due to autocorrelation inherent in time-series data. Time-series which are highly autocorrelated appear correlated with one another, even when there is no underlying causal relationship (see **Figure 1**). This leads to high false-positive rates of inferred interactions, particularly for correlation-based inference methods (Kurtz et al., 2015; Weiss et al., 2016; Coenen and Weitz, 2018; Carr et al., 2019; Hirano and Takemoto, 2019; Mainali et al., 2019; Thurman et al., 2019).

Model-based inference methods are built from dynamical models of microbial community ecology. As such, temporal variation and structure is incorporated into any model-based inference framework, accounting for potentially difficult

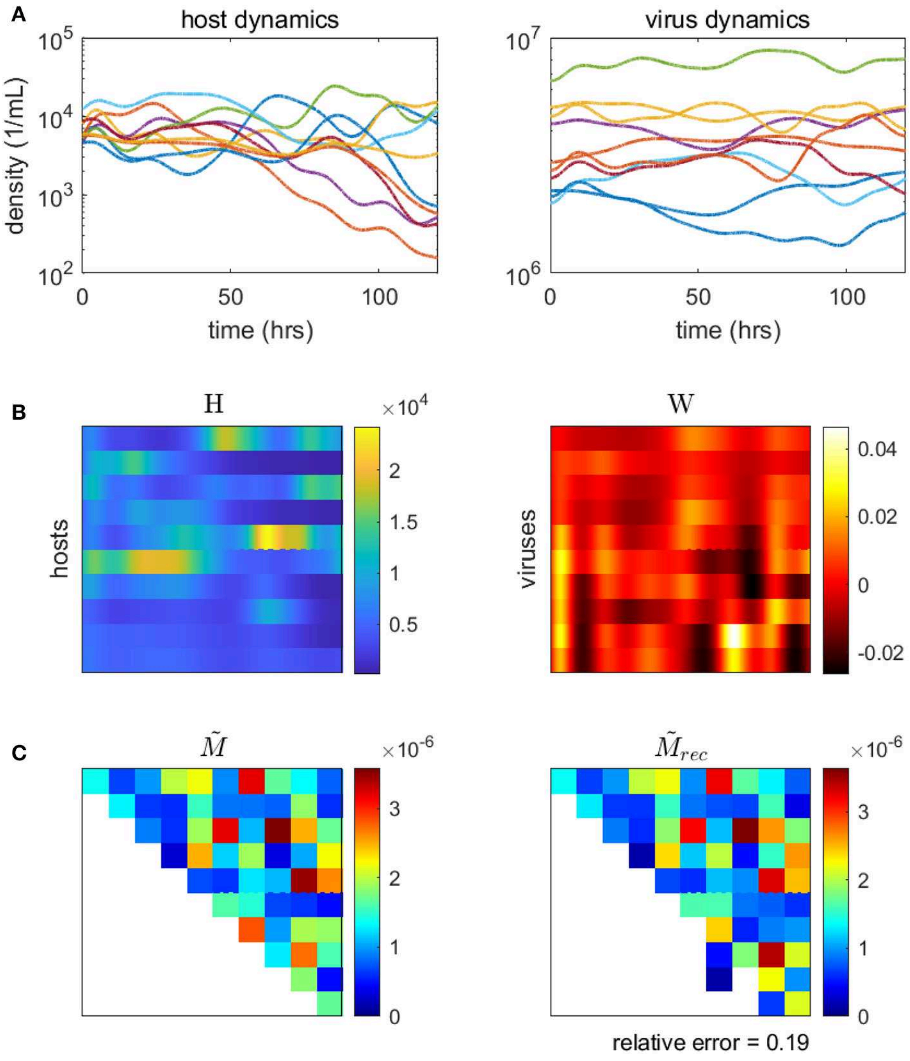


FIGURE 6 | Inferring the microbe-virus infection network from time-series data for a 10 by 10 synthetic microbe-virus community. **(A)** Simulated host (left) and virus (right) densities over time. **(B)** Host densities (left, H) and transformed virus differences (right, W), for input into the objective function (Equation 20). **(C)** The original "ground-truth" interaction network (left) and the reconstructed network (right). In the interaction matrix, the rows denote hosts, the columns represent viruses, and the colors denote the scaled intensity of interactions (where white denotes no interaction).

statistical properties, such as autocorrelation. Model-based inference has been shown to perform favorably in *in silico* studies (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Major challenges remain for implementing model-based inference in practice, including requirements of high time-resolution data and a detailed understanding of the biological and ecological mechanisms at play in order to correctly specify the underlying model. Futhermore, evaluating accuracy of inferred networks remains dificult, in part because diferent networks can produce similar patterns of ecological dynamics (Cao et al., 2017). Despite challenges, model-based inference has shown potential to accurately infer interaction networks in a computationally efficient and scalable manner (see one such application in Stein et al., 2013).

Here, we demonstrate the use of a model-based inference method on a synthetic microbial community with viruses (methods and code adapted from Jover et al., 2016). We use a synthetic community so that the inferred network can be compared to the original, “ground-truth” network. Using our model for microbe-virus ecological dynamics (Equation 17), we simulate population time-series of the community over the course of several days. We sample the simulated time-series to use as data inputs into the minimization problem (Equation 20), from which we estimate the weighted microbe-virus infection network $\tilde{M}$. Simulated time-series, data inputs, original and reconstructed networks are shown in **Figure 6**). As shown, the reconstructed network closely resembles the original, with only minor quantitative differences (i.e., in the strengths of the interactions). We caution that the choice (and parameterization) of ecological dynamics is critical to developing a model-based approach, for alternative examples see Mounier et al. (2008), Stein et al. (2013), Fisher and Mehta (2014), Marino et al. (2014), Dam et al. (2016), Jover et al. (2016), Ovaskainen et al. (2017), Xiao et al. (2017), Faust et al. (2018), and Venturelli et al. (2018).

4. CONCLUSION

The aim of this primer was to integrate analytic advances together to serve practical aims, so that they can be transferred for analysis of other high resolution temporal data sets. Conducting high-resolution temporal analyses to understand microbial community dynamics has become more feasible in recent years with continued advances in sequence technology. Accordingly, specific statistical considerations should be taken into account as a precursor for microbiome analysis. In this primer, we summarized challenges in analyzing time-series data and present examples which synthesize practical steps to manage these challenges. For further reading on the topics addressed here, we recommend: normalizations and log-ratios (Egozcue et al., 2003; Silverman et al., 2017), distance calculations (Willis and Martin, 2018), clustering (Kurtz et al., 2015; Martin-Platero et al., 2018), statistical ordination (Morton et al., 2017; Ren et al., 2017), regression (Martin et al., 2019), vector autoregression (Opgen-Rhein and Strimmer, 2007), periodicity detection (Ernst and Bar-Joseph, 2006), general best practices (Holmes and Huber, 2019), and an in-depth review of multivariate data analysis (Buttigieg and Ramette, 2014). For inferring interactions from time-series, model-based inference approaches have significant potential (Mounier et al., 2008; Stein et al., 2013; Fisher and Mehta, 2014; Marino et al., 2014; Dam et al., 2016; Jover et al., 2016; Ovaskainen et al., 2017; Xiao et al., 2017; Faust et al., 2018; Venturelli et al., 2018). Although correlation-based methods have been widely used for inferring interactions, recent literature suggests that correlation-based methods are poor indicators of interaction (Weiss et al., 2016; Coenen and Weitz, 2018; Carr et al., 2019; Hirano and Takemoto, 2019; Mainali et al., 2019; Thurman et al., 2019). Other model-free methods, such as Granger causality (Mainali et al., 2019) and cross-convergent mapping (Sugihara et al., 2012), may be useful alternatives for inference although care should be taken that data do not violate the methods’ assumptions (McCracken and Weigel, 2014; Baskerville and Cobey, 2017). In closing, we hope that the consolidated methods and workflows in both R and Matlab help researchers from multiple disciplines advance the quantitative *in situ* study of microbial communities.

DATA AVAILABILITY STATEMENT

For the 18S rRNA gene-based survey, data originated from Hu et al. (2018). The raw sequence data can also be found under SRA BioProject PRJNA393172. Code to process this 18S rRNA tag-sequencing data can be found at https://github.com/shu251/18Sdiversity_diel and quality checked reads and final OTU table used for downstream data analysis is available (10.5281/zenodo.1243295), as well as in the GitHub https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

AUTHOR CONTRIBUTIONS

AC, SH, EL, DM, and JW conceptualized the work. SH provided the data for analysis. AC, DM, and JW designed the methods and analyses. SH and DM wrote the code for the clustering and periodicity tutorials. AC wrote the code for the inference tutorial. AC, SH, EL, DM, and JW co-wrote the manuscript. All authors approved the manuscript.

FUNDING

This work was supported by the Simons Foundation (SCOPE award ID 329108) and the National Science Foundation (NSF Bio Oc 1829636).

ACKNOWLEDGMENTS

We thank Dave Caron for helpful feedback and multiple reviewers for their feedback on this manuscript.

REFERENCES

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. *J. Control Decis.* 5, 42–60. doi: 10.1080/23307706.2017.1397554

Aitchison, J. (1983). The statistical analysis of compositional data. *J. Int. Assoc. Math. Geol.* 44, 139–177.

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

Aylward, F. O., Boeuf, D., Mende, D. R., Wood-Charlson, E. M., Vislova, A., Eppley, J. M., et al. (2017). Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc. Natl. Acad. Sci. U.S.A.* 114, 11446–11451. doi: 10.1073/pnas.1714821114

Aylward, F. O., Eppley, J. M., Smith, J. M., Chavez, F. P., Scholin, C. A., and DeLong, E. F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5443–5448. doi: 10.1073/pnas.1502883112

Bairey, E., Kelsic, E. D., and Kishony, R. (2016). High-order species interactions shape ecosystem diversity. *Nat. Commun.* 7:12285. doi: 10.1038/ncomms12285

Baskerville, E. B., and Cobey, S. (2017). Does influenza drive absolute humidity? *Proc. Natl. Acad. Sci. U.S.A.* 114, E2270–E2271. doi: 10.1073/pnas.1700369114

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

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., et al. (2005). Defining operational taxonomic units using dna barcode data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1935–1943. doi: 10.1098/rstb.2005.1725

Borcard, D., and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Modell.* 153, 51–68. doi: 10.1016/S0304-3800(01)00501-4

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecol. Monogr.* 27, 325–349.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3932–3937.

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* 3, 1–27. doi: 10.1080/03610917408548446

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. doi: 10.1038/ismej.2017.119

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

Caron, D. A. (2013). Towards a molecular taxonomy for protists: benefits, risks, and applications in plankton ecology. *J. Eukaryot. Microbiol.* 60, 407–413. doi: 10.1111/jeu.12044

Caron, D. A., and Hu, S. K. (2019). Are we overestimating protistan diversity in nature? *Trends Microbiol.* 27, 197–205. doi: 10.1016/j.tim.2018.10.009

Carr, A., Diener, C., Baliga, N. S., and Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *ISME J.* 13, 2674–2655. doi: 10.1038/s41396-019-0459-z

Charvet, S., Vincent, W. F., and Lovejoy, C. (2014). Effects of light and prey availability on Arctic freshwater protist communities examined by high-throughput DNA and RNA sequencing. *FEMS Microbiol. Ecol.* 88, 550–564. doi: 10.1111/1574-6941.12324

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using z score transformation. *J. Mol. Diagn.* 5, 73–81. doi: 10.1016/S1525-1578(10)60455-2

Coenen, A. R., and Weitz, J. S. (2018). Limitations of correlation-based inference in complex virus-microbe communities. *mSystems* 3:e00084–18. doi: 10.1128/mSystems.00084-18

Conneely, K. N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of *p* values for multiple correlated tests. *Am. J. Hum. Genet.* 81, 1158–1168. doi: 10.1086/522036

Dam, P., Fonseca, L. L., Konstantinidis, K. T., and Voit, E. O. (2016). Dynamic models of the complex microbial metapopulation of lake mendota. *NPJ Syst. Biol. Appl.* 2:16007. doi: 10.1038/npjsba.2016.7

Diamond, S., and Boyd, S. (2016). CVXPY: a python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* 17, 1–5.

Egozcue, J. J., Pawlowsky-Glahn, V., Figueras, G., and Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9:968. doi: 10.1038/ismej.2014.195

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

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

Finkel, Z. V., Beardall, J., Flynn, K. J., Quigg, A., Rees, T. A. V., and Raven, J. A. (2009). Phytoplankton in a changing world: cell size and elemental stoichiometry. *J. Plankton Res.* 32, 119–137. doi: 10.1093/plankt/fbp098

Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* 9:e0102451. doi: 10.1371/journal.pone.0102451

Gerea, M., Queimaliños, C., and Unrein, F. (2019). Grazing impact and prey selectivity of picoplanktonic cells by mixotrophic flagellates in oligotrophic lakes. *Hydrobiologia* 831, 5–21. doi: 10.1007/s10750-018-3610-3

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* 67, 850–857. doi: 10.1016/j.jclinepi.2014.03.012

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

Gower, J. C., and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *J. classif.* 3, 5–48.

Grant, M., and Boyd, S. (2008). "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences*, eds V. Blondel, S. Boyd, and H. Kimura (Springer-Verlag Limited), 95–110. Available online at: http://stanford.edu/boyd/graph_dcp.html

Grant, M., and Boyd, S. (2014). *CVX: Matlab Software for Disciplined Convex Programming, Version 2.1*. Available online at: http://cvxr.com/cvx

Grilli, J., Barabás, G., Michalska-Smith, M. J., and Allesina, S. (2017). Higher-order interactions stabilize dynamics in competitive network models. *Nature* 548, 210–213. doi: 10.1038/nature23273

Gülagiz, F. K., and Sahin, S. (2017). Comparison of hierarchical and non-hierarchical clustering algorithms. *Int. J. Comput. Eng. Inform. Technol.* 9:6.

Hein, M., Pedersen, M. F., and Sand-Jensen, K. (1995). Size-dependent nitrogen uptake in micro-and macroalgae. *Mar. Ecol. Prog. Ser.* 118, 247–253. doi: 10.3354/meps118247

Hirano, H., and Takemoto, K. (2019). Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics* 20:329. doi: 10.1186/s12859-019-2915-1

Holmes, S., and Huber, W. (2019). *Modern Statistics for Modern Biology*. Cambridge, UK: Cambridge University Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24:417. doi: 10.1037/h0071325

Hu, S. K., Campbell, V., Connell, P., Gellene, A. G., Liu, Z., Terrado, R., et al. (2016). Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific. *FEMS Microbiol. Ecol.* 92:fiw050. doi: 10.1093/femsec/fiw050

Hu, S. K., Connell, P. E., Mesrop, L. Y., and Caron, D. A. (2018). A hard day's night: diel shifts in microbial eukaryotic activity in the north pacific subtropical gyre. *Front. Mar. Sci.* 5:351. doi: 10.3389/fmars.2018.00351

Hu, S. K., Liu, Z., Lie, A. A. Y., Countway, P. D., Kim, D. Y., Jones, A. C., et al. (2015). Estimating protistan diversity using high-throughput sequencing. *J. Eukaryot. Microbiol.* 62, 688–693. doi: 10.1111/jeu.12217

Hughes, M. E., Abruzzi, K. C., Allada, R., Anafi, R., Arpat, A. B., Asher, G., et al. (2017). Guidelines for genome-scale analysis of biological rhythms. *J. Biol. Rhythms* 32, 380–393. doi: 10.1177/0748730417728663

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/annotation/3d8a6578-ce56-45aa-bc71-05078355b851

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytol.* 11, 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Jover, L. F., Cortez, M. H., and Weitz, J. S. (2013). Mechanisms of multi-strain coexistence in host-phage systems with nested infection networks. *J. Theor. Biol.* 332, 65–77. doi: 10.1016/j.jtbi.2013.04.011

Jover, L. F., Romberg, J., and Weitz, J. S. (2016). Inferring phage-bacteria infection networks from time-series data. *R. Soc. Open Sci.* 3:160654. doi: 10.1098/rsos.160654

Karl, D. M. (2002). Hidden in a sea of microbes. *Nature* 415, 590–591. doi: 10.1038/415590b

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T.-K., Lua, R. C., Wilkins, A. D., et al. (2014). Single nucleotide variations: biological impact and theoretical interpretation. *Prot. Sci.* 23, 1650–1666. doi: 10.1002/pro.2552

Kaufman, L., and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, Vol. 344. New York, NY: John Wiley & Sons.

Kavanaugh, M. T., Hales, B., Saraceno, M., Spitz, Y. H., White, A. E., and Letelier, R. M. (2014). Hierarchical and dynamic seascapes: a quantitative framework for scaling pelagic biogeochemistry and ecology. *Prog. Oceanogr.* 120, 291–304. doi: 10.1016/j.pocean.2013.10.013

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S. (2014). "Dbscan: past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* (Nicosia), 232–238.

Kim, M., Morrison, M., and Yu, Z. (2011). Evaluation of different partial 16s rrna gene sequence regions for phylogenetic analysis of microbiomes. *J. Microbiol. Methods* 84, 81–87. doi: 10.1016/j.mimet.2010.10.020

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

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B* 361, 1929–1940. doi: 10.1098/rstb.2006.1920

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Korytowski, D. A., and Smith, H. L. (2017). Persistence in phage-bacteria communities with nested and one-to-one infection networks. *Discrete Contin. Dyn. Syst. B* 22, 859–875. doi: 10.3934/dcdsb.2017043

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* 37, 233–243. doi: 10.1002/aic.690370209

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115–129. doi: 10.1007/BF02289694

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* 7, 813–819. doi: 10.1038/nmeth.1499

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

Liao, T. W. (2005). Clustering of time series data–a survey. *Pattern Recogn.* 38, 1857–1874. doi: 10.1016/j.patcog.2005.01.025

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). "Understanding of internal clustering validation measures," in *Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10* (Washington, DC: IEEE Computer Society), 911–916. doi: 10.1109/ICDM.2010.35

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lozupone, C., 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

Luo, E., Aylward, F. O., Mende, D. R., and DeLong, E. F. (2017). Bacteriophage distributions and temporal variability in the ocean's interior. *mBio* 8:e01903–17. doi: 10.1128/mBio.01903-17

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420

Mainali, K., Bewick, S., Vecchio-Pagan, B., Karig, D., and Fagan, W. F. (2019). Detecting interaction networks in the human microbiome with conditional granger causality. *PLoS Comput. Biol.* 15:e1007037. doi: 10.1371/journal.pcbi.1007037

Mangan, N. M., Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Inferring biological networks by sparse identication of nonlinear dynamics. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* 2, 52–63.

Mangan, N. M., Kutz, J. N., Brunton, S. L., and Proctor, J. L. (2017). Model selection for dynamical systems via sparse regression and information criteria. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* 473, 20170009.

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

Mann, M. E., and Lees, J. M. (1996). Robust estimation of background noise and signal detection in climatic time series. *Clim. Change* 33, 409–445. doi: 10.1007/BF00142586

Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. (2014). Mathematical modeling of primary succession of murine intestinal microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 111, 439–444. doi: 10.1073/pnas.1311322111

Martin, B. D., Witten, D., and Willis, A. D. (2019). Modeling microbial abundances and dysbiosis with beta-binomial regression. *arXiv* 1902.02776.

Martin-Platero, A. M., Cleary, B., Kauffman, K., Preheim, S. P., McGillicuddy, D. J., Alm, E. J., et al. (2018). High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat. Commun.* 9:266. doi: 10.1038/s41467-017-02571-4

McCracken, J. M., and Weigel, R. (2014). Convergent cross-mapping and pairwise asymmetric inference. *Phys. Rev. E* 90:062903. doi: 10.1103/PhysRevE.90.062903

McKie-Krisberg, Z. M., Gast, R. J., and Sanders, R. W. (2015). Physiological responses of three species of antarctic mixotrophic phytoflagellates to changes in light and dissolved nutrients. *Microb. Ecol.* 70, 21–29. doi: 10.1007/s00248-014-0543-x

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *bioRxiv*. doi: 10.7554/eLife.46923.027

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

Mende, D. R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nat. Methods* 10:881. doi: 10.1038/nmeth.2575

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., et al. (2017). Balance trees reveal microbial niche differentiation. *mSystems* 2:e00162–16. doi: 10.1128/mSystems.00162-16

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A.-S., Hélias, A., et al. (2008). Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* 74, 172–181. doi: 10.1128/AEM.01338-07

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. Compstat Lectures, Vienna: Physika Verlag.

Noble, W. S. (2009). How does multiple testing correction work? *Nat. Biotechnol.* 27:1135. doi: 10.1038/nbt1209-1135

Nygaard, K., and Tobiesen, A. (1993). Bacterivory in algae: a survival strategy during nutrient limitation. *Limnol. Oceanogr.* 38, 273–279. doi: 10.4319/lo.1993.38.2.0273

Opgen-Rhein, R., and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics* 8:S3. doi: 10.1186/1471-2105-8-S2-S3

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin, R., Schuster, S. C., et al. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345, 207–212. doi: 10.1126/science.1252476

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E., et al. (2017). How are species interactions structured in species-rich communities? a new method for analysing time-series data. *Proc. Biol. Sci*. 284, 20170768. doi: 10.1098/rspb.2017.0768

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. doi: 10.1038/nmeth.2658

Poretsky, R. S., Hewson, I., Sun, S., Allen, A. E., Zehr, J. P., and Moran, M. A. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the north pacific subtropical gyre. *Environ. Microbiol*. 11, 1358–1375. doi: 10.1111/j.1462-2920.2008.01863.x

Ren, B., Bacallado, S., Favaro, S., Holmes, S., and Trippa, L. (2017). Bayesian nonparametric ordination for the analysis of microbial communities. *J. Am. Stat. Assoc*. 112, 1430–1442. doi: 10.1080/01621459.2017.1288631

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., et al. (2015). Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proc. Natl. Acad. Sci. U.S.A*. 112, 8008–8012. doi: 10.1073/pnas.1424279112

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537:689. doi: 10.1038/nature19366

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 6:e21887. doi: 10.7554/eLife.21887

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol*. 9:e1003388. doi: 10.1371/journal.pcbi.1003388

Stevens, J. R., Al Masud, A., and Suyundikov, A. (2017). A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLoS ONE* 12:e0176124. doi: 10.1371/journal.pone.0176124

Storch, D., and Šizling, A. L. (2008). The concept of taxon invariance in ecology: Do diversity patterns vary with changes in taxonomic resolution? *Folia Geobotanica*.

Streiner, D. L. (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity-whether and how to correct for many statistical tests. *Am. J. Clin. Nutr*. 102, 721–728. doi: 10.3945/ajcn.115.113548

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., et al. (2012). Detecting causality in complex ecosystems. *Science* 338, 496–500. doi: 10.1126/science.1227079

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

Thaben, P. F., and Westermark, P. O. (2014). Detecting rhythms in time series with rain. *J. Biol. Rhythms* 29, 391–400. doi: 10.1177/0748730414553029

Thamatrakoln, K., Talmy, D., Haramaty, L., Maniscalco, C., Latham, J. R., Knowles, B., et al. (2019). Light regulation of coccolithophore host-virus interactions. New Phytol. 221, 1289–1302.

Thurman, L. L., Barner, A. K., Garcia, T. S., and Chestnut, T. (2019). Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography* 42, 1658–1670. doi: 10.1111/ecog.04360

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

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, T., Mavrommatis, K., Kyrpides, N. C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res*. 43, 6761–6771. doi: 10.1093/nar/gkv657

Venturelli, O. S., Carr, A. C., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., et al. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol*. 14:e8157. doi: 10.15252/msb.20178157

Vincenzi, S., Crivelli, A. J., Munch, S., Skaug, H. J., and Mangel, M. (2016). Trade-offs between accuracy and interpretability in von bertalanffy random-effects models of growth. *Ecol. Appl*. 26, 1535–1552.

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. 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

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., et al. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J*. 10:2557. doi: 10.1038/ismej.2016.45

Willis, A. D. (2019). Rigorous Statistical Methods for Rigorous Microbiome Science. *MSystems* 4, e00117–19. doi: 10.1128/mSystems.00117-19

Willis, A. D., and Martin, B. D. (2018). Divnet: estimating diversity in networked communities. *bioRxiv*. doi: 10.1101/305045

Wilson, S. T., Aylward, F. O., Ribalet, F., Barone, B., Casey, J. R., Connell, P. E., et al. (2017). Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium crocosphaera. *Nat. Microbiol*. 2:17118. doi: 10.1038/nmicrobiol.2017.118

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., and Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. *Nat. Commun*. 8:2042. doi: 10.1038/s41467-017-02090-2

Xu, D., Li, R., Hu, C., Sun, P., Jiao, N., and Warren, A. (2017). Microbial eukaryote diversity and activity in the water column of the south china sea based on DNA and RNA high throughput sequencing. *Front. Microbiol*. 8:1121. doi: 10.3389/fmicb.2017.01121

Yang, R., and Su, Z. (2010). Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics* 26, i168–i174. doi: 10.1093/bioinformatics/btq189

Yang, R., Zhang, C., and Su, Z. (2011). LSPR: an integrated periodicity detection algorithm for unevenly sampled temporal microarray data. *Bioinformatics* 27, 1023–1025. doi: 10.1093/bioinformatics/btr041

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A., and Elshahed, M. S. (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rRNA gene-based environmental surveys. *Appl. Environ. Microbiol*. 75:5227. doi: 10.1128/AEM.00592-09

Zobitz, J. M., Desai, A. R., Moore, D. J. P., and Chadwick, M. A. (2011). A primer for data assimi lation with ecological models using markov chain monte carlo (mcmc). *Oecologia*.

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

Cascabel: A Scalable and Versatile Amplicon Sequence Data Analysis Pipeline Delivering Reproducible and Documented Results

Alejandro Abdala Asbun[1], *Marc A. Besseling*[1], *Sergio Balzano*[1†], *Judith D. L. van Bleijswijk*[1], *Harry J. Witte*[1], *Laura Villanueva*[1,2] *and Julia C. Engelmann*[1*]

[1] Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands, [2] Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Utrecht, Netherlands

OPEN ACCESS

Edited by:
Jens Stoye,
Bielefeld University, Germany

Reviewed by:
Sebastian Menke,
University of Ulm, Germany
Sebastian Jünemann,
Bielefeld University, Germany

***Correspondence:**
Julia C. Engelmann
julia.engelmann@nioz.nl

†Present address:
Sergio Balzano,
Stazione Zoologica Anton Dohrn,
Instituto Nazionale di Biologia,
Ecologia e Biotecnologie Marine,
Naples, Italy

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

Received: *02 August 2019*
Accepted: *05 October 2020*
Published: *20 November 2020*

Citation:
Abdala Asbun A, Besseling MA, Balzano S, van Bleijswijk JDL, Witte HJ, Villanueva L and Engelmann JC (2020) Cascabel: A Scalable and Versatile Amplicon Sequence Data Analysis Pipeline Delivering Reproducible and Documented Results. Front. Genet. 11:489357. doi: 10.3389/fgene.2020.489357

Marker gene sequencing of the rRNA operon (16S, 18S, ITS) or cytochrome c oxidase I (CO1) is a popular means to assess microbial communities of the environment, microbiomes associated with plants and animals, as well as communities of multicellular organisms *via* environmental DNA sequencing. Since this technique is based on sequencing a single gene, or even only parts of a single gene rather than the entire genome, the number of reads needed per sample to assess the microbial community structure is lower than that required for metagenome sequencing. This makes marker gene sequencing affordable to nearly any laboratory. Despite the relative ease and cost-efficiency of data generation, analyzing the resulting sequence data requires computational skills that may go beyond the standard repertoire of a current molecular biologist/ecologist. We have developed *Cascabel*, a scalable, flexible, and easy-to-use amplicon sequence data analysis pipeline, which uses Snakemake and a combination of existing and newly developed solutions for its computational steps. *Cascabel* takes the raw data as input and delivers a table of operational taxonomic units (OTUs) or Amplicon Sequence Variants (ASVs) in BIOM and text format and representative sequences. *Cascabel* is a highly versatile software that allows users to customize several steps of the pipeline, such as selecting from a set of OTU clustering methods or performing ASV analysis. In addition, we designed *Cascabel* to run in any linux/unix computing environment from desktop computers to computing servers making use of parallel processing if possible. The analyses and results are fully reproducible and documented in an HTML and optional pdf report. *Cascabel* is freely available at Github: https://github.com/AlejandroAb/CASCABEL.

Keywords: amplicon sequencing, 16S/18S rRNA gene, Illumina, community profiling, microbiome, pipeline, snakemake

1. INTRODUCTION

High-throughput sequencing of an omnipresent marker gene, such as the gene coding for the small subunit of the ribosomal RNA (16S for prokaryotes or 18S for eukaryotes) is a cost-efficient means for community profiling that is affordable for nearly every lab. On current sequencing platforms, up to hundreds of samples can be combined (multiplexed) in a single sequencing run, decreasing the sequencing costs per sample tremendously, and generating massive amounts of data. Not surprisingly, community compositions based on DNA analyses have been generated from most of the habitats on earth, including the human body (Human Microbiome Project Consortium, 2012), the open ocean (Sunagawa et al., 2015), deep sea (Sogin et al., 2006), and intracellular symbionts (Balzano et al., 2015). Moreover, sequencing a marker gene like cytochrome c oxidase I (CO1) or mitochondrial 12S in environmental DNA also allows to track larger multicellular organisms, for example fish in the sea (Hänfling et al., 2016; van Bleijswijk et al., 2020). Amplicon sequencing can also be used to investigate active microbial communities based on ribosomal RNA abundance instead of the rRNA gene locus (Massana et al., 2015; Forster et al., 2016). Typically, a short fragment of 100–600 nucleotides of the marker gene is amplified by PCR from the DNA extract or cDNA generated from the rRNA extract of the community, and then sequenced by high throughput sequencing. During sequence analysis, sequences are often grouped in Operational Taxonomic Units (OTUs) following one of two main strategies: *de novo* or *closed-reference* OTU picking (Westcott and Schloss, 2015). With closed-reference OTU picking, sequences are assigned to a sequence from a reference database given an identity threshold. Sequences which are not similar enough to any sequence in the database are excluded from downstream analyses.

De novo OTU picking clusters reads sharing a predefined sequence identity, commonly 97%, yielding approximately species resolution considering the entire 16S rRNA gene (about 1,500 nt) (Stackebrandt and Goebel, 1994). Although widely used, its application to short read sequencing data has been criticized because the individual variable regions of the 16S rRNA gene have quite different taxonomic resolution for different groups of organisms, such that it is impossible to find a general cutoff of sequence identity which would reliably distinguish species (Johnson et al., 2019). For eukaryotes, the situation is similar with respect to the taxonomic resolution at a given sequence identity threshold of the 18S gene, as this can vary even within the same taxonomic group. For example, it has been shown that within diatoms, *Nitzschia* and *Thalassiosira* species can be easily separated based on the diversity of their 18S rRNA gene (Hoppenrath et al., 2007; Rimet et al., 2011) whereas distinct *Pseudo-nitzschia* and *Chaetoceros* species share identical 18S rRNA gene sequences (Amato et al., 2007; Balzano et al., 2017).

In response to the criticism of OTUs, alternative methods have been developed which model sequencing errors to estimate the true biological sequence. DADA2 (Callahan et al., 2016) and deblur (Amir et al., 2017) cluster reads such that the clusters are consistent with the error model, while Minimum entropy Decomposition (MED) (Eren et al., 2015) and Swarm (Mahé et al., 2015) assume that sequence errors occur randomly and they use this assumption and abundance information of unique sequences to cluster them into supposedly biological entities. To set them apart from OTUs, the term "Amplicon Sequence Variant (ASV)" has been coined for results from denoising algorithms, such as DADA2 and Deblur, while MED uses "oligotypes" and Swarm "swarms" for their clusters. All of these approaches do not require setting a sequence identity threshold, and the resolution is determined by the data, which seems to better reflect the true state of nature (Caruso et al., 2019). However, OTU methods are still widely being used and might deliver useful insights for applications where lower taxonomic resolution is sufficient.

While the experimental part of community profiling studies is fairly simple (DNA extraction, PCR), the current bottleneck is the computational analysis of the (potentially massive) sequence data. For scientists with little background in bioinformatics, the amount of data and complexity of data analysis can be overwhelming. Popular software solutions for the individual steps from raw sequence data to an OTU or ASV table, e.g., QIIME (Caporaso et al., 2010b), mothur (Schloss et al., 2009), and DADA2 (Callahan et al., 2016), are not necessarily straightforward to use. The software package mothur (Schloss et al., 2009), which comes with its own computational environment, and the QIIME framework (Caporaso et al., 2010b) both require the ability to work on the command line. Analyzing multiple sequencing libraries quickly becomes tedious for users not proficient in implementing bash (or any other programming language) scripts which chain the individual steps and allow parallel processing. The ASV analysis tool DADA2 (Callahan et al., 2016) comes as an R package, which also requires some scripting skills. While web servers for microbial community data analysis like SILVAngs (Quast et al., 2013) and MG-RAST (Glass et al., 2010), NGTax2 (Poncheewin et al., 2019) and SLIM (Dufresne et al., 2019) are easy to use, they are inherently inflexible and also limited in throughput. QIIME2 (Bolyen et al., 2018) has command line and graphical user interface (GUI) modes of operation and offers even a larger choice of algorithms for data analysis than the original QIIME, including statistical analyses of the resulting community profiles. The GUI has limited functionality though and might not be a convenient solution for analyzing many samples. The same holds for recently developed GUIs like BTW (Morais et al., 2018) and SEED2 (Vetrovský et al., 2018) which run under Microsoft Windows, and PipeCraft (Anslan et al., 2017) which provides a GUI running on Linux systems. For example, none of these three provide an ASV analysis method. More recently developed pipelines which run on the command line focus on usability with minimal bioinformatic skills, but allowing higher throughput than a webserver. These recent pipelines frequently chain existing tools to make them more accessible, but often at the cost of flexibility due to fixed parameter settings, e.g. BMPOS (Pylro et al., 2016), BTW (Morais et al., 2018), and MetaAmp (Dong et al., 2017), or fixed reference databases, like PEMA (Zafeiropoulos et al., 2020). Others miss essential functionality which requires additional tools to make them useful, e.g., PEMA (Zafeiropoulos et al., 2020), and iMAP (Buza et al., 2019) do not provide demultiplexing of sequence libraries.

None of the easy-to-use tools cited above allow the analysis of sequence read pairs which are not overlapping (DADA2 supports non-overlapping reads, but requires R skills). Although most often amplicons are designed and sequenced such that forward and reverse reads overlap and can be merged into one continuous sequence, for example the primer pair 515F (GTGCCAGCMGCCGCGGTAA), 926R (CCGYCAATTYMTTTRAGTTT) amplifies bacterial and archaeal 16S as well as eukaryotic 18S regions (Needham and Fuhrman, 2016). This makes it a cost-efficient approach if both prokaryotic and eukaryotic communities are of interest. With marine environmental samples, the primers produce an amplicon of on average 411 nucleotides originating from prokaryotic sequences and an amplicon of on average 585 nucleotides derived from eukaryotic sequences. Forward and reverse reads from the longer eukaryotic amplicon typically do not overlap sufficiently (current maximum read length of an Illumina MiSeq is 2 × 300 nucleotides) to merge both reads, especially when amplicon length varies between species, and low quality and adapter sequences are trimmed from the reads. The MeFit pipeline (Parikh et al., 2016) can merge forward and reverse reads with N characters, but it is merely a merging and filtering tool, not a complete amplicon data analysis pipeline. A complete workflow to carry on with this kind of analysis is currently not available, to the best of our knowledge. *Cascabel* allows to "stitch" together the forward and reverse non- or not sufficiently overlapping read pairs with any character and continues with the analysis. In existing pipelines, these sequences would be discarded.

Moreover, most of the existing tools do not have documentation functions to guarantee reproducibility and facilitate communicating which software tools, their versions and parameter settings were used. We could identify only one tool, iMAP (Buza et al., 2019), developed at the same time as *Cascabel*, which can generate a report of the analysis. However, iMAP requires editing and adapting bash scripts before it can be run and is therefore not very user-friendly.

Therefore, we anticipated a need for a pipeline which combines the flexibility and scalability provided by using bioinformatic tools on the command line with the ease of using interactive web servers for analyzing and interpreting amplicon sequencing data.

Moreover, issues with reproducibility of research findings have made data provenance an important aspect of data analysis and scientific journals start to require documentation of data provenance for submitted manuscripts (www.nature.com, 2019). Not all pipelines are transparent enough to trace back all the exact steps taken by underlying software and their versions used within the pipeline. With interactive webservers it is often impossible to reproduce an analysis because once the webserver is updated, previous versions are no longer accessible, or, even worse, the webserver changes without the user noticing. Even if versions are documented and previous versions are available, it is the responsibility of the user to actively record all parameter settings in an unambiguous way, which is an error-prone endeavor. Also, this information cannot be recovered at a later time point, and wrong documentations are likely to go unnoticed.

One of the main strengths of *Cascabel* is that all analyses (runs) performed are completely documented and reproducible. All scripts of *Cascabel* are located within the project folder, and together with the Snakemake and configuration file, every run of the pipeline is completely reproducible at any time. All software versions used are documented in the run reports (in HTML and pdf format). The code of all *Cascabel* scripts is open source, although we use some third-party modules which are not open source, e.g., UCLUST, USEARCH (Edgar, 2010).

We here provide *Cascabel*, a Snakemake (Köster and Rahmann, 2012) pipeline for the analysis of community marker gene sequence data which is easy to use for people with little bioinformatics background, and both flexible and powerful enough to be attractive for people with bioinformatics training. *Cascabel* supports large sample and sequencing library throughput as well as parallel computing on personal computers and computing servers. Moreover, the results are summarized in an HTML and pdf report, and all input and output files, tools, parameters and their versions are documented, rendering the analyses fully reproducible.

2. IMPLEMENTATION

Our pipeline makes use of the workflow management engine Snakemake (Köster and Rahmann, 2012), which scales from personal workstations to computer clusters. *Cascabel* consists of a set of "rules" which specify the input, the action to perform on the input (executed by a bash/python/R/java script), and the output. The user defines *via* a configuration file (called "config file" from now on) in yaml format, how these "rules" are chained to perform amplicon sequence data analysis from the raw data to the final OTU or ASV table. *Cascabel* saves the OTU or ASV table in BIOM and text format to allow further analysis and interpretation with statistical or visualization tools. For most of the rules, *Cascabel* provides several alternative algorithms or tools and allows passing arguments *via* the config file to the algorithm being used. In addition, rules can be skipped, and the pipeline can be entered and exited at every step. This makes *Cascabel* very flexible and highly customizable. Moreover, the pipeline is easily extendable and amendable to personal needs, allowing for example the analysis of any marker gene sequence data.

In addition to securing data provenance, *Cascabel* has a suite of rules (modules) which are not readily available or straightforward for a non-bioinformatician to implement with existing amplicon sequence data analysis tools. The first one is a custom dereplication rule for very large data sets based on VSEARCH (Rognes et al., 2016), which, depending on the duplication level of the sequence data, can up to double the number of reads which can be dereplicated on a given system. Second, *Cascabel* supports the analysis of non-overlapping read pairs arising from long amplicons. *Cascabel* can "stitch" these reads together with any desired sequence of characters, e.g., one or several N and then proceeds with OTU or ASV analysis. We recommend using RDP for taxonomic classification as this k-mer based method is not affected by additional N characters.

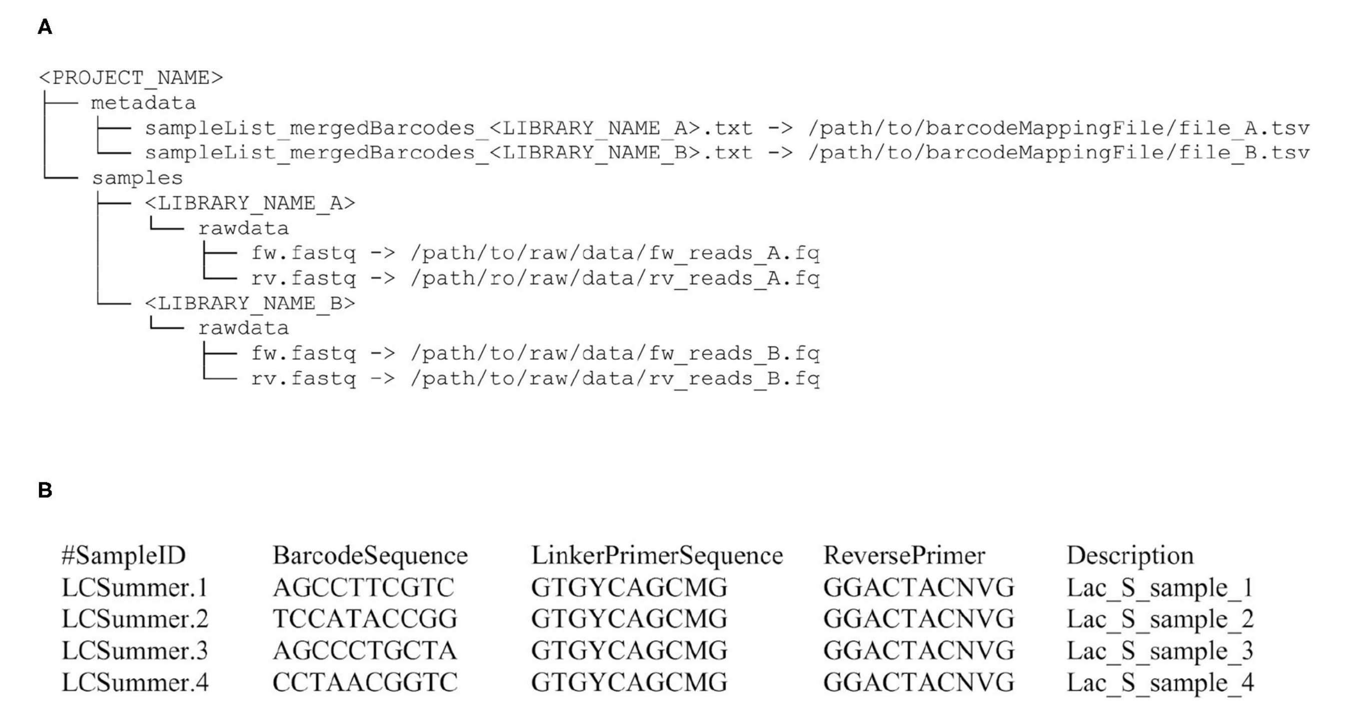
Furthermore, *Cascabel* can generate data files for submission to a public sequence read archive. Demultiplexed fastq files can be generated with barcodes, or barcodes and primers/adapters removed, ready for submission. Another unique feature of *Cascabel* is that the user can determine the level of interaction with the pipeline. In interactive mode, the user is informed about the results of individual rules and can amend parameters during runtime, while in non-interactive mode the pipeline will proceed according to the parameter settings in the config file. We outline the individual steps performed by *Cascabel* below.

Running *Cascabel* requires the raw fastq sequence data files, a mapping file indicating which sample carries which barcode if the data should be demultiplexed, and optionally sample metadata (e.g., geographic coordinates of the sampling stations, physical, chemical, or biological properties), and the config file specifying the tools and parameters used for running the pipeline. When working with one sequencing library, users can pass file paths to the raw data and metadata directly in the config file. When working with multiple libraries, users can choose between listing the input file names in a text file and referring to this file in the config file, or using the helper script *initSample.sh*. This script is run for each library to initialize the folder structure expected by *Cascabel*. With all three options, the folder structure will look like the one illustrated in **Figure 1A**.

We provide example config files with default parameters for double- and single barcoded paired-end reads for OTU and ASV analysis on the github page of *Cascabel* (https://github.com/AlejandroAb/CASCABEL). However, we strongly advise to make informed choices about parameter settings matching the individual needs of the experiment and data set. With the files in place, *Cascabel* is started with a one-line command on the terminal. Snakemake takes care of executing the rules in a computationally efficient manner, making optimal use of available resources, e.g., distributing jobs over several nodes. **Figure 2** provides an overview of the workflow of *Cascabel*. In **Table 1** we summarize the options and methods provided for the individual steps of the analysis performed by *Cascabel*.

Cascabel has an interactive and a non-interactive mode. In interactive mode, several modules have a check-point which needs to be passed to continue with the analysis. If the check fails (e.g., if too many FastQC (Andrews, 2010) quality modules failed or the number of sequences assigned to sample barcodes is too low), the pipeline stops and the user has to decide to continue, change parameters and continue, or exit the pipeline. If parameters were changed interactively, the new ones are documented in the reports. The interactive mode is useful in the explorative data analysis stage, while the non-interactive mode is suitable for running large batches of data and evaluating the results later.

The first step of *Cascabel* consists of checking the validity of the input files including the barcode mapping file and the config file. *Cascabel* supports single-end as well as paired-end sequence data as input from one or multiple samples per input file. Barcodes for demultiplexing samples can be situated at the beginning of one or both of the reads. The barcode sequences are read from the barcode mapping file, which is exemplified in **Figure 1B**. **Supplementary Datasheets 1**, **2** contain sample config files, which were used to generate the reports provided in **Supplementary Datasheets 3–5**. After having validated the input files, *Cascabel* proceeds with analyzing sequence data quality with FastQC (Andrews, 2010). In interactive mode, *Cascabel* will stop if more than a specified number of quality check modules failed.



A

```
<PROJECT_NAME>
├── metadata
│   ├── sampleList_mergedBarcodes_<LIBRARY_NAME_A>.txt -> /path/to/barcodeMappingFile/file_A.tsv
│   └── sampleList_mergedBarcodes_<LIBRARY_NAME_B>.txt -> /path/to/barcodeMappingFile/file_B.tsv
└── samples
    ├── <LIBRARY_NAME_A>
    │   └── rawdata
    │       ├── fw.fastq -> /path/to/raw/data/fw_reads_A.fq
    │       └── rv.fastq -> /path/ro/raw/data/rv_reads_A.fq
    └── <LIBRARY_NAME_B>
        └── rawdata
            ├── fw.fastq -> /path/to/raw/data/fw_reads_B.fq
            └── rv.fastq -> /path/to/raw/data/rv_reads_B.fq
```

B

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Description
LCSummer.1	AGCCTTCGTC	GTGYCAGCMG	GGACTACNVG	Lac_S_sample_1
LCSummer.2	TCCATACCGG	GTGYCAGCMG	GGACTACNVG	Lac_S_sample_2
LCSummer.3	AGCCCTGCTA	GTGYCAGCMG	GGACTACNVG	Lac_S_sample_3
LCSummer.4	CCTAACGGTC	GTGYCAGCMG	GGACTACNVG	Lac_S_sample_4

FIGURE 1 | Input file structure for *Cascabel*. **(A)** This input file structure is generated from the file paths provided in the config file when the dataset consists of a single sequencing library. For multiple libraries, it is created from a text file specifying the individual libraries or by the helper script initSample.sh. **(B)** Example of a barcode mapping file for four samples. Barcode and primer sequences are listed in 5′-3′ direction and have been abbreviated.

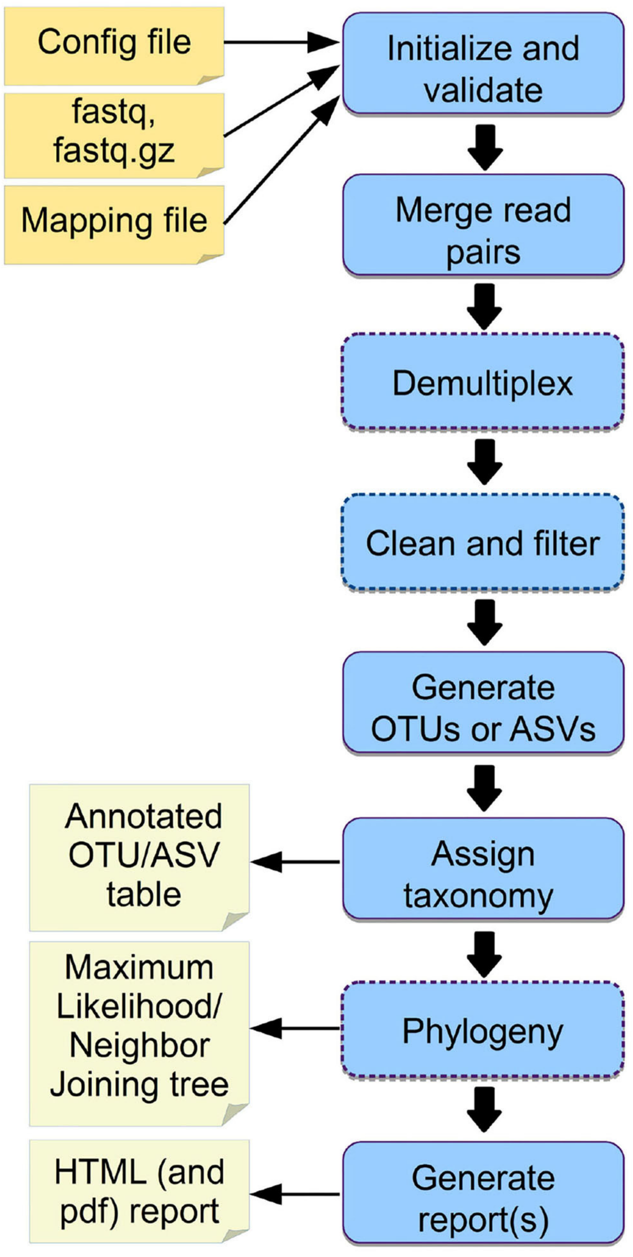


FIGURE 2 | Overview of *Cascabel*. The workflow indicates input files (config file, sequence data in fastq format, barcode mapping file), mandatory and optional steps of the pipeline (blue boxes) as well as the main output files. The boxes of optional steps have dashed borders. "Clean and filter" refers to removing primers/adapters and chimeras. **Table 1** shows a detailed summary of the steps, available tools and output files.

TABLE 1 | Outline of the steps performed by *Cascabel*. "Script(s)" refers to *Cascabel* scripts in bash, java or R.

Step	Tools/Algorithms	Output
Initialize structure	Script	Project folder and file structure
Quality Control	FastQC (Andrews, 2010)	FastQC report
Merge reads	PEAR (Zhang et al., 2014)	Merged (assembled) sequences
Demultiplex	QIIME (Caporaso et al., 2010b), scripts	Sequences assigned to samples in one file and per sample
Align vs. reference	Mothur (Schloss et al., 2009)	Aligned sequences
Remove chimeras	usearch61 (Edgar, 2010), Uchime_denovo and uchime_ref (VSEARCH) (Rognes et al., 2016)	Chimera-free sequences
Remove adapters	Cutadapt (Martin, 2011)	Adapter-free sequences
Size filter	Script	Filtered sequences
Dereplicate	VSEARCH	Dereplicated sequences
Generate OTUs	Mothur (Schloss et al., 2009), prefix/suffix (Caporaso et al., 2010b), CD-HIT (Li and Godzik, 2006), SUMACLUST (Kopylova et al., 2016), Swarm (Mahé et al., 2015), UCLUST (Edgar, 2010), trie (Caporaso et al., 2010b) sortmerna (Kopylova et al., 2012)	OTU table
Pick representatives (OTUs)	Random, longest, most_abundant, first	Fasta file with representative sequences
Generate ASVs	DADA2 (Callahan et al., 2016)	ASV table
Assign taxonomy OTUs	QIIME [BLAST (Altschul et al., 1990), UCLUST, RDP (Wang et al., 2007)], blastn (BLAST+) (Camacho et al., 2009), VSEARCH	Taxonomic assignments for each OTU
Assign taxonomy ASVs	RDP	Taxonomic assignments for each ASV
Generate OTU table	QIIME, scripts	Annotated OTU table
Generate ASV table	DADA2	Annotated ASV table
Alignment	Pynast (Caporaso et al., 2010a), mafft (Katoh and Standley, 2013), infernal (Nawrocki and Eddy, 2013), clustalw (Larkin et al., 2007), muscle (Edgar, 2004)	Multiple sequence alignment
Make tree	Muscle, clustalw, raxml (Stamatakis, 2006), fasttree (Price et al., 2009)	Phylogenetic tree
Report	Scripts, Krona (Ondov et al., 2011)	HTML, pdf report, Krona charts

Next, read pairs are assembled with PEAR (Zhang et al., 2014) and the quality of the assembled reads is again assessed with FastQC. *Cascabel* also offers an "unpaired" workflow for paired-end sequence data with non-overlapping reads. For this kind of data, *Cascabel* merges the forward and reverse read with an "N" or any other character, and assigns taxonomy using the RDP classifier, which, due to using a k-mer approach, is not impacted by this procedure (Jeraldo et al., 2014).

If the library contains sequences from several samples, they are demultiplexed based on the barcode sequences provided in the barcode mapping file. To do so, *Cascabel* makes use of QIIME (Caporaso et al., 2010b) and a custom R script to (optionally) allow sequence errors in the barcodes.

Demultiplexed data can also be stored in individual fastq files for further use outside the pipeline, e.g., for submitting data to public repositories. Optionally, *Cascabel* will align sequence reads against a reference sequence database to remove off-target reads and facilitate removing sequence adapters or primers or both. Adapter and primer sequences can be trimmed off with Cutadapt (Martin, 2011). Then, *Cascabel* generates a histogram of sequence lengths. In interactive mode, *Cascabel* shows the frequency of occurrence of each of the read lengths on the terminal and allows to change the minimum and maximum sequence length provided in the config file. The library report contains a smoothed histogram of the sequence lengths to validate the choice of the minimum and maximum sequence length (**Figure 3A**). Optionally, *Cascabel* identifies and removes chimeras either *de novo* based on sequence abundance or searching against the gold database provided by QIIME with the usearch61 algorithm (Edgar, 2010). The user can also provide different databases, such as SILVA (Quast et al., 2013) or PR2 (Guillou et al., 2013) to search for chimeras. Assembled and potentially filtered sequence reads from all samples are then concatenated into one fasta file. *Cascabel* generates a histogram to visualize the number of reads per sample for each of the libraries to assess whether the sequences are evenly spread across the samples (**Figure 3B**). Furthermore, the reports for each of the libraries contain a plot of the number and percentages of raw, assembled, demultiplexed and length filtered sequences (**Figure 3C**).

When working with large datasets, a dereplication step which collapses identical sequences into one representative sequence can drastically reduce computation time. *Cascabel* provides a custom rule based on VSEARCH (Rognes et al., 2016). *Cascabel*'s dereplication rule splits the data in two chunks and dereplicates them individually first, which, depending on how many duplicate sequences there are in the dataset, up to doubles the number of reads which can be dereplicated with the available memory. Then, the two chunks of dereplicated reads are merged and again dereplicated. To generate an OTU table, the dereplications are traced back by *Cascabel*.

Cascabel provides a range of popular methods to generate OTUs with or without a reference sequence database [Swarm (Mahé et al., 2015), sortmerna (Kopylova et al., 2012), mothur (Schloss et al., 2009), trie (Caporaso et al., 2010b), UCLUST/UCLUST_REF/USEARCH/USEARCH_REF (Edgar, 2010), prefix/suffix (Caporaso et al., 2010b), CD-HIT (Li and Godzik, 2006), and SUMACLUST (Kopylova et al., 2016)], some of these are executed by QIIME.

Then, representative sequences are chosen for each OTU (with options: random, longest, most_abundant, first) (Caporaso et al., 2010b). OTU and representative sequence picking methods provided by *Cascabel* are listed in **Table 1**. From the abundances of the OTU sequences within each of the samples, *Cascabel* creates an OTU abundance table. The OTUs can further be grouped at higher taxonomic levels depending on the desired resolution. An overview of the folder structure and main output files generated by *Cascabel* is given in **Supplementary Datasheet 6**.

Alternatively, *Cascabel* can perform Amplicon Sequence Variant (ASV) analysis with DADA2 (Callahan et al., 2016) for paired-end sequence data. In this case, *Cascabel* takes the demultiplexed fastq files and passes them to various R scripts which run sequence filtering, ASV identification, chimera detection and taxonomic assignment with DADA2. The main output of the ASV analysis are an ASV count table and ASV representative sequences. An example config file for an ASV analysis can be found in **Supplementary Datasheet 2**, and the ASV report for this analysis is shown in **Supplementary Datasheet 5**. The main output files of the ASV analysis are shown in **Supplementary Datasheet 7**.

Cascabel can process sequence data from any marker gene. *Cascabel* comes with taxonomic mapping files for 16S rRNA and 18S rRNA gene sequences from SILVA v132 (Quast et al., 2013), but the user can always choose to make use of a different public or a custom reference sequence database. *Cascabel* provides three different approaches to assign taxonomy to the representative sequences: VSEARCH, which performs global alignment of the target sequences against the reference database; BLAST, making use of BLAST+ (Camacho et al., 2009); QIIME, with methods BLAST (Altschul et al., 1990), UCLUST or the RDP classifier. Alternatively, any other public or custom database can be used for taxonomic annotation. If taxonomy is assigned with VSEARCH or BLAST, the user can choose to assign the sequences to the lowest common ancestor (LCA) with the stampa approach (https://github.com/frederic-mahe/stampa).

Subsequently, the user can opt to remove singletons, align representative sequences, filter the alignment and make a phylogenetic tree. To align representative sequences, *Cascabel* offers pynast (Caporaso et al., 2010a), mafft (Katoh and Standley, 2013), infernal (Nawrocki and Eddy, 2013), clustalw (Larkin et al., 2007), and muscle (Edgar, 2004). A phylogenetic tree can be generated with muscle, clustalw, raxml (Stamatakis, 2006) and fasttree (Price et al., 2009) (**Table 1**).

The last rule of *Cascabel* (the "target" rule) generates HTML and optional pdf reports with documentation, figures and tables summarizing the results of individual rules, as well as all software versions and parameter settings used. If more than one library was analyzed, there will be a report for each library as well as a report summarizing all libraries (otu_report or asv_report). Among other graphics, the otu_report shows the percentages and the total number of reads after filtering ("combined reads"), dereplicated reads, OTUs, OTUs assigned to a taxonomic level, OTUs excluding singletons ("no singletons"), and assigned OTUs excluding singletons (**Figure 3D**). The asv_report shows similar information in a table. **Supplementary Datasheet 3** shows an example library report, **Supplementary Datasheet 4** an otu_report, and **Supplementary Datasheet 5** an asv_report. In addition, *Cascabel* generates an interactive Krona chart (Ondov et al., 2011) for the run which displays community composition for individual samples or the complete data set. The Krona chart shows the taxonomic assignments in an interactive HTML document composed of a multi-layered pie-chart and the user can zoom in and browse these different levels. An example is shown in **Figure 3E**.

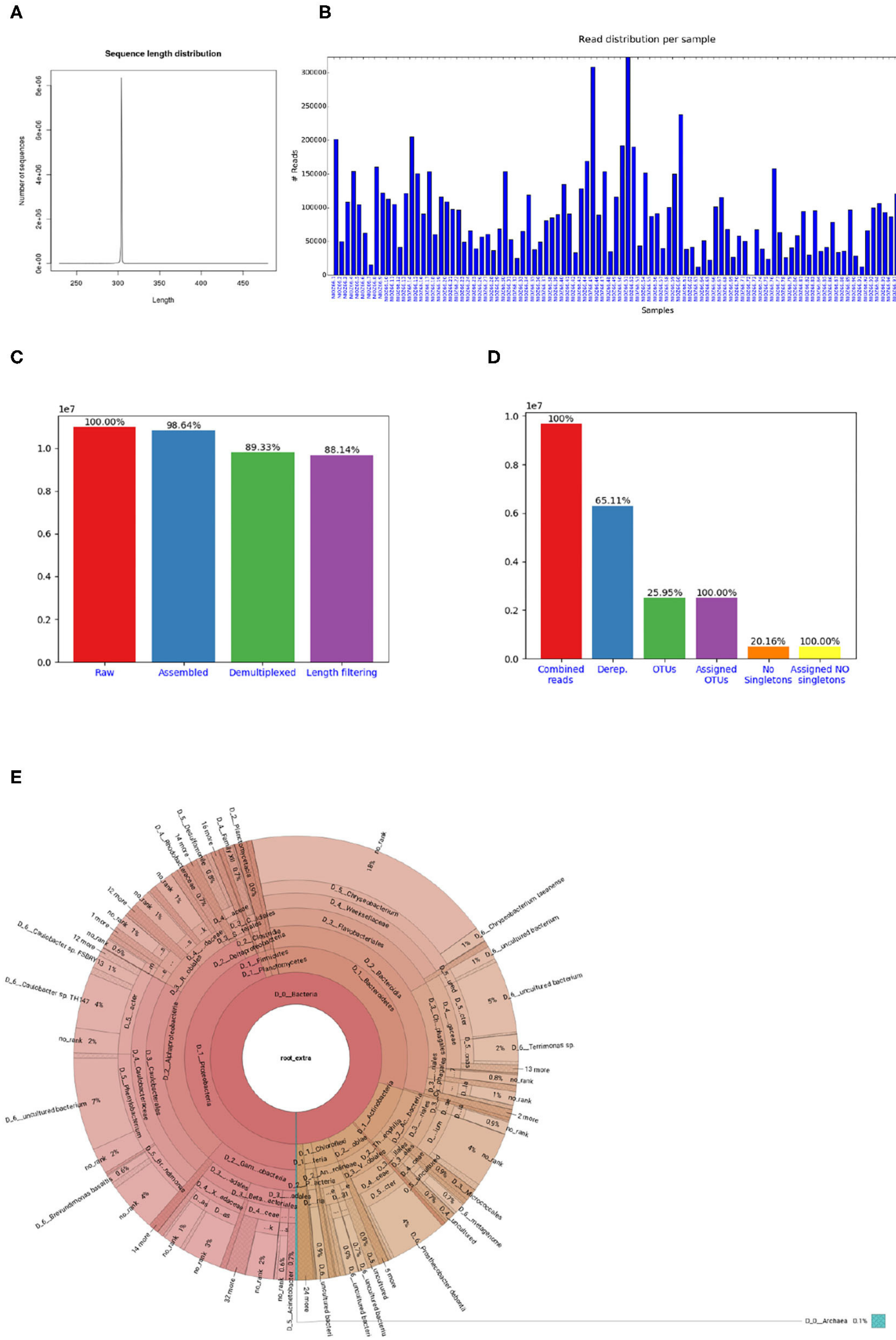


FIGURE 3 | Figures shown in *Cascabel* reports. **(A)** Smoothed sequence length distribution after merging reads, for one library. The plot is meant to help making a sensible choice for sequence length filtering. **(B)** Number of sequences per sample. This histogram is part of the OTU report (including all libraries). **(C)** Number of sequences after individual pre-processing steps. "Assembled" refers to the number of raw read pairs which could be merged based on their overlap. "Demultiplexed"

(Continued)

FIGURE 3 | refers to the number of raw reads which could be assembled and assigned to a sample, and "Length filtering" indicates the number of raw reads passing the previous and the sequence length criteria. This plot is part of the library report. **(D)** Number of sequences after individual steps after potentially combining several libraries (total number of reads) and generating OTUs. "Derep." indicates the number of dereplicated reads and their percentage relative to the total combined reads. "OTUs" is the total number of OTUs and the percentage is relative to the number of combined reads. "Assigned OTUs" is the number and percentage of OTUs with a taxonomic assignment. "No singletons" refers to the number and percentage of OTUs excluding singleton OTUs, and "Assigned NO singletons" is the number and percentage of singleton-free OTUs with a taxonomic assignation. The plot is part of the OTU report. **(E)** Krona chart for one sample. The krona charts are interactive and can be viewed with a web browser. Colors indicate the taxonomic groups to which the OTU was assigned. Each ring of the pie chart represents a different taxonomic level. An example of a full library report is shown in **Supplementary Datasheet 3**, and an OTU report is provided in **Supplementary Datasheet 4**.

A unique feature of *Cascabel* is its native handling of multiple analyses on the same dataset. Snakemake will not re-run a rule if the output file of that rule already exists, unless `--forcerun` is used or the input file has a more recent date than the existing output file. This avoids unintentional over-writing of existing results, but also renders it impossible to keep results of multiple analyses on the same data in the same project. To allow multiple analyses within the same project, we implemented *Cascabel* with a "Run" parameter. Whenever the user changes the "Run" parameter, a new analysis will be performed (except for quality control on the raw data) and the results saved in a different "Run" folder. Each run has its own reports and is therefore fully documented and reproducible. Confusion about parameter settings for a specific analysis cannot happen.

The "Run" parameter is also useful to analyse data from primers which generate multiple fragments with different lengths. The data can then be analyzed with individual runs for each expected fragment length. For example, when using primers which amplify both bacterial/archaeal 16S and eukaryotic 18S sequences, albeit with different length, one run can target the shorter fragment and a second run the longer fragment. If the longer fragment generates non-overlapping read pairs, these can be analyzed with the "unpaired" workflow as indicated in the config file.

To facilitate comparing different taxonomy assignment approaches, the user can perform taxonomic assignments for the same run using different methods and the results will be saved in individual "taxonomy" folders. When starting a new taxonomic assignment, the existing OTU representative sequences are used so no processing time is wasted by performing the same upstream rules several times.

The user can make use of all intermediate files generated by individual rules, and most importantly the OTU or ASV table and representative sequences for follow-up analyses. To save disk space, the user can also opt to have *Cascabel* remove temporary files at the end of the analyses. For many rules, the user can pass additional parameters to the command or tool at hand using the "extra_params" parameter in the config file.

3. RESULTS

3.1. *Cascabel* Example Analysis of a 16S Dataset

To demonstrate the functionality of *Cascabel*, we applied it to 16S rRNA gene amplicon data generated from water column samples taken from Lake Chala. *Cascabel* offers two routes of analysis: OTU and ASV analysis. Some rules apply to both routes, others only to one of them. This is indicated in the header of the rule in the config file by either "BOTH_WF," "OTU_WF," or "ASV_WF." If choosing "OTU_WF," for example, the "ASV_WF" rules and their parameter settings are ignored. We chose 'ANALYSIS_TYPE: "OTU"' here. After validating the input files, *Cascabel* proceeded with analyzing sequence data quality with FastQC (Andrews, 2010). In interactive mode, *Cascabel* will stop if more than a specified number of quality check modules failed, in non-interactive mode it will proceed. Next, we assembled read pairs, which is mandatory for paired-end data (rule "pear"). If the amplicon is so long that the forward and reverse read do not overlap, *Cascabel* can be run using the rule "UNPAIRED DATA WORK FLOW" and setting "UNPAIRED_DATA_PIPELINE" to true. After forward and reverse read assembly or merging, the quality of the assembled reads was again assessed with FastQC. Then, the sequencing library was demultiplexed based on the barcode sequences provided in the barcode mapping file (rules "write_dmx_files" and "extract_barcodes"). This step is optional to allow processing already demultiplexed data. We demultiplexed the Lake Chala data based on a sample barcode of 12 nucleotides at the beginning of the forward and the reverse read, using the "barcode_paired_stitched" configuration which merges the barcode sequence of the forward read with the barcode sequence of the reverse read. Barcode sequences were provided in a barcode mapping file, such as exemplified in **Figure 1B**. Individual barcodes were designed such that they have a nucleotide difference of at least three, however, we allowed only two mismatches in the merged barcode of 24 nucleotides to assign reads to samples to avoid false positive assignments due to sequencing errors. The demultiplexing rule can also save demultiplexed data in individual fastq files for further use outside the pipeline, e.g., for submitting data to public repositories, by setting the "create_fastq_files" parameter to "T" (true). During demultiplexing, technical sequences, such as primers can also be removed, and we did so for the Lake Chala data (primers are indicated in the config file, **Supplementary Datasheet 1**).

After demultiplexing, sequence chimeras can be removed based on a reference database, e.g., the gold database, and/or *de novo* based on sequence abundance, but we set this rule ("search_chimera") to false. The next step is to filter out sequences with unexpected length. To facilitate setting length thresholds, *Cascabel* generates a smoothed histogram of observed sequence lengths, which is shown in the library report (**Figure 3A**). In interactive mode, *Cascabel* also shows the frequency of each of the read lengths on the terminal, and allows to change the minimum and maximum sequence length provided in the config file on the command line. For the analyzed example data, we filtered out sequences whose length differed more than 10 nucleotides from the average sequence length. Next, we dereplicated sequences which were identical over the full sequence length (rule "dereplicate"). This step is optional, but recommended to decrease the runtime of OTU clustering and avoid memory issues with very large datasets.

For OTU clustering, we chose UCLUST with a similarity threshold of 0.97, resulting in roughly 2.5 million OTUs (rule "cluster_OTUs"). We selected the longest sequence of an OTU as representative sequence to be used for taxonomic assignment of the OTU (rule "pick_representatives"). Then we used VSEARCH to assign taxonomy to the representative sequences based on the SILVA database (SILVA version 132, rule "assign_taxonomy"). From the abundances of the OTU sequences within each of the samples, *Cascabel* creates an OTU abundance table in BIOM and plain text format (rule "make_otu_table"). The OTUs were also summarized at different taxonomic levels making use of the rule "summarize_taxa." Subsequently, we removed singletons (rule "filter_otu"), aligned representative sequences ("align_rep_seqs"), filtered the alignment ("filter_alignment"), and made a phylogenetic tree ("make_tree"). Removing singletons reduced the number of OTUs in the analyzed dataset to roughly $500,000$. To align representative sequences, we used pynast and fasttree to generate a phylogenetic tree. Finally, *Cascabel* generated a Krona chart, and HTML and pdf reports of the analysis, documenting all software and parameters used. **Supplementary Datasheet 1** contains the config file with parameter settings for the analysis described above.

The total runtime of *Cascabel* using the OTU workflow on the Lake Chala dataset was 14.5 h. Currently, we experience a bottleneck in the runtime of the barcode error correction, which took 10.25 h on this large dataset and will be improved in future versions of *Cascabel*. We also ran the ASV workflow on the same data (config file shown in **Supplementary Datasheet 2**), which took 13.2 h in total, with again barcode error correction being the most time-consuming step (10.25 h). Running *Cascabel* assigning reads with perfectly matching barcodes only would take 4.25 h for the OTU workflow and 2.95 h for the ASV workflow.

3.2. Analyses of Mock Datasets

We evaluated the results of the individual runs in terms of the number of genera identified correctly (true positives), the number of genera missed (false negatives) and the number of genera identified which were not part of the mock community (false positives). We evaluated all runs with respect to true and false positives, and show the individual true and false positive genera for a selection of the runs which we performed on the mock datasets. The selection included at least one run using UCLUST, one using Swarm and one ASV run with DADA2. We also varied Swarm parameters, reference databases, clustering thresholds, and chimera detection, but evaluating all possible combinations of parameters would not be feasible. An overview of the runs performed and the evaluation in terms of true and false positives, false negatives, precision, recall and F1 statistic (harmonic mean of precision and recall), is shown in **Supplementary Datasheet 8**. While on the 16S mock dataset, all of UCLUST, swarm and DADA2 had a very good recall rate of 0.95, the OTU/ASV clustering methods had different numbers of false positive predictions. DADA2 had the lowest number of false positives (1), followed by Swarm (14–27) and UCLUST (21–25). Therefore, DADA2 performed best in terms of precision (0.95) and F1 statistic (0.95). The 18S mock dataset was more challenging than the 16S dataset for all combinations of methods tested. UCLUST with a similarity threshold of 0.97 and VSEARCH for taxonomy assignment using SILVA v138 performed best in terms of recall (0.92). However, DADA2 performed best concerning precision (1.0) and the harmonic mean of precision and recall (0.8). On the ITS dataset, the best performance was shown by Swarm using $d = 2$ and VSEARCH for taxonomy assignation, with an F1 statistic of 0.92. This run also showed the best recall (0.89). The highest precision was achieved by UCLUST (1.0), however, with a very low F1 statistic (0.19). The performance of DADA2 was lower than the one of Swarm, with an F1 statistic of 0.8. Thus, we observed substantial differences between different methods and parameter settings, and there was no one setting that would perform best on all three datasets. On the contrary, the best results were obtained with different methods and parameter settings for different marker genes. These results confirm that it is important to have a flexible pipeline to adapt it to the needs of the dataset at hand, but also that it is important to include a mock community ideally in every sequencing run that is performed to allow making informed choices about method and parameter selection. We also compared the different clustering algorithms in terms of runtime. The 16S mock community consisted of 207,197 paired-end reads of 300 nucleotides each, considerably smaller than the Lake Chala dataset, and therefore the analyses were much faster. The analysis with UCLUST (config_4.yaml) took 43 min and 18 s, of which 38 min and 42 s were spent on searching chimeras *de novo*. Swarm needed 5 min and 16 s for a run including chimera search against a reference database (config_12.yaml), searching chimeras took 1 min and 24 s of the total time. A DADA2 run (config_14.yaml) needed 8 min and 52 s including DADA2's own chimera search method.

4. DISCUSSION

Cascabel has been developed at the Royal Netherlands Institute for Sea Research (NIOZ) to facilitate, unify and easily track data provenance of amplicon sequence data analyses. Apparent advantages of using this pipeline compared to custom scripts are that the individual steps of the pipeline have been tested by many members of the community at the NIOZ who are experienced in amplicon sequencing data analyses (van Bleijswijk et al., 2015; Balzano et al., 2018; Besseling et al., 2019; Klunder et al., 2019), and therefore should contain fewer mistakes than scripts that were written for a specific analysis by one person. Moreover, community knowledge and experience have created a workflow which is probably more comprehensive and powerful than one that was created by a single person. In addition, the availability of the pipeline has facilitated comparing and integrating research results from different data sets generated at the NIOZ because scientists can agree on certain settings and reference database versions and the pipeline guarantees that the analyses are performed in the same way. Because *Cascabel* keeps track of data provenance, documenting the process of analyzing the data to generate results, it also facilitates preparing research manuscripts. While most of the scientific journals request the raw sequencing data to be submitted to a public repository for many years already, also reporting data provenance becomes more important. The journal "Nature," for example, requires authors to make materials, data, code, and associated protocols available (www.nature.com, 2019). *Cascabel* facilitates providing data, code and protocols. Public sequence repositories often require the raw data to be submitted per sample, but sample demultiplexing typically takes place after merging read pairs such that the raw data cannot be recovered. Therefore, *Cascabel* demultiplexes the raw data in parallel to the analyses such that it is ready for public data repository deposition. The code of Cascabel is open source and all analyses are protocolled in the reports and config file, complying with the rules for reproducible computational research described by Sandve et al. (2013).

DNA sequencing technology, algorithms and analysis approaches are constantly evolving. It is logical that pipelines lag behind with the most recent developments because it takes time to test and integrate new modules. Because *Cascabel* is a Snakemake workflow, it is flexible and easy to extend to encompass more or alternative rules. We are constantly working on extending the range of applications and making new methods available. For the sake of consistency, we deliberately keep older methods to allow users to compare runs using their familiar algorithm with newer algorithms and to compare or integrate new data with data generated previously.

The task of generating biological meaningful microbial community profiles from amplicon sequence data is far from trivial, and we believe that there is not one best strategy for data analysis. Based on the environment investigated and the scientific question, desired taxonomic resolution may differ. Therefore, we do not want to promote any optimal settings of the tools used by *Cascabel*. We do, however, provide some guidance by making pre-configured config files available, but advise any user to check them carefully and modify them to their needs. Our analyses on public mock datasets have shown that the optimal method may depend on the marker gene and the dataset at hand. Therefore, we advise users to evaluate their favorite configurations for an analysis on mock datasets and ideally include a mock community in their own sequencing projects.

Cascabel provides reference databases for taxonomy assignment and chimera detection, but the user can always supply a different database and specify that in the config file. Moreover, *Cascabel* is not limited to Illumina sequence data that we used for demonstration purposes, but can handle sequence data from other technologies which produce short reads from amplicons as well (e.g., Ion Torrent). With some minor modifications, *Cascabel* can even be used to analyze long read amplicon sequence data.

Galaxy (Afgan et al., 2018) is a user-friendly web-based alternative to *Cascabel* which offers interfaces to VSEARCH and mothur executables. Having a medium-sized user group at the institute, we did not want to overload a public server and setting up and maintaining our own server would also need resources that we preferred to allocate to the development of a workflow for which we have full control and flexibility. With *Cascabel* being invoked from the command line, the user can make use of the full potential that Snakemake has to offer, e.g., `--prioritize` to force the execution of specific rules prior to others when distributing tasks across computing resources, `--until` to run the pipeline up to a specific rule, `--summary`, which shows the rules executed so far and `--dag` which shows the rules executed and the ones yet to be done in a directed acyclic graph. Moreover, we consider *Cascabel*'s report an essential element to move forward in terms of user-friendly data provenance and reproducibility.

We have presented *Cascabel*, an open source pipeline to analyze amplicon sequence data based on the workflow engine Snakemake. The pipeline can be easily installed using Anaconda or Miniconda, comes with documentation, a wiki, and a test dataset on github and can be executed by users with basic command line skills. At the same time, *Cascabel* is flexible, offering alternative options for most of the steps and supporting custom reference databases, and can easily be modified and extended by users with computational skills. Moreover, all analyses performed with *Cascabel* are fully documented and reproducible. We believe that *Cascabel* will prove to be useful to scientist who need more flexibility and throughput than provided by tools based on web servers, but do not want to or cannot generate their own command-line based workflow.

5. METHODS

5.1. Sampling, DNA Extraction, and Sequencing of Example Dataset

Suspended particulate matter (SPM) was collected from the water column of Lake Chala, a lake situated on the border of Kenya and Tanzania, east of Mount Kilimanjaro in Africa, from September 2013 to May 2014 from a total of 111 samples as described in van Bree et al. (2018). DNA was extracted from 1/32 section of the filters on which SPM was collected by using the PowerSoil DNA extraction kit (Mo Bio Laboratories, Carlsbad, CA, USA).

The V4 region of the 16S rRNA gene was amplified with the primers forward:
515F: GTGYCAGCMGCCGCGGTAA (Parada et al., 2016) and reverse:
806RB: GGACTACNVGGGTWTCTAAT (Apprill et al., 2015). We made use of 12 nucleotide Golay barcodes at the beginning of the forward and the reverse read. Paired-end sequencing of 250 nt was performed on an Illumina MiSeq instrument (Illumina, San Diego, CA) using the Truseq DNA nano LT kit for library preparation and V3 sequencing chemistry at the sequencing facility of the University of Utrecht (USEQ), the Netherlands. The dataset contains a total of 10, 979, 168 paired-end sequence reads. The data is publicly available at NCBI, BioProject PRJNA526242. Sample and run identifiers of the samples used are listed in **Supplementary Datasheet 9**.

5.2. Analysis of Example Dataset

Starting from the config file template for paired-end sequencing data (config.otu.double_bc.yaml, provided on github: https://github.com/AlejandroAb/CASCABEL), we supplied the file paths to the raw sequence data (fastq files) in the "GENERAL PARAMETERS SECTION" (subsection "INPUT FILES"). Note that fastq files can also be provided as gzipped files, then in the "INPUT TYPE" section of the general parameter section, the parameter *gzip_input* needs to be set to "T" (True). The barcode mapping file was passed to *Cascabel via* the "metadata" parameter in the subsection "INPUT FILES" of the config file. In the "GENERAL PARAMETERS SECTION," we chose a project name ("CascabelTest") and set the "RUN" parameter to "report_test," *Cascabel* then used these names to generate a project folder and a run folder. All settings and parameters chosen to analyze the example data set are documented in the config files (**Supplementary Datasheets 1**, **2**) and the reports (**Supplementary Datasheets 3–5**). The reports also contain software versions of third-party tools incorporated in *Cascabel*.

5.3. Analyses of Mock Community Datasets

To show the flexibility and assess the performance of running *Cascabel* with different methods and parameter settings, we analyzed three published mock community datasets with multiple *Cascabel* runs. We chose one dataset consisting of 16S rRNA data, one of 18S rRNA data and one of ITS sequences. The 16S rRNA dataset is part of the public resource project for bioinformatics benchmark data, Mockrobiota (Bokulich et al., 2016). The mock community is composed of 20 evenly distributed bacterial strains as described in Gohl et al. (2016). For the ITS marker gene, we used a dataset of Bakker (2018), composed of 19 fungal species with staggered abundances, intended to mimic the abundance distribution of natural microbial communities. Finally, for the 18S rRNA marker gene, we selected a mock community composed of 12 algal species across five major divisions of eukaryotic microalgae (Bradley et al., 2016). More information about the selected datasets, sample accessions, links to the rawdata and different parameters used to run *Cascabel* can be found in **Supplementary Datasheet 8**. The config files of the individual runs are available in **Supplementary Datasheet 10** and at *Cascabel's* test data repository (https://github.com/AlejandroAb/CASCABEL-Test/tree/master/mock_analysis). Fastq files were downloaded from the European Nucleotide Archive (ENA) using an in-house download tool (https://github.com/AlejandroAb/ENA-downloader-tool).

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the SRA of NCBI, BioProject PRJNA526242.

AUTHOR CONTRIBUTIONS

AA implemented the pipeline, with contributions from JE. MB, LV, SB, and HW tested the pipeline. AA, JB, and JE designed the pipeline, with contributions from LV. JE wrote the manuscript. All authors contributed to and approved the final version of the manuscript.

FUNDING

LV and AA were supported by the Soehngen Institute of Anaerobic Microbiology (SIAM) Gravitation grant (024.002.002) of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organization for Scientific Research (NWO).

ACKNOWLEDGMENTS

We thank Hans Malschaert for IT support and the bioinformatics community at the NIOZ for comments and suggestions about the pipeline. We acknowledge Sanne Vreugdenhil and Maartje Brouwer for technical and analytical support. Fieldwork with collection of the studied sample materials was carried out with permission from the government of Kenya through permit 13/001/11C to Dirk Verschuren. In accordance with National Environmental Management Authority regulations in the context of the Nagoya Protocol, DNA extracts of the analyzed suspended-particulate samples are archived at the National Museums of Kenya (NMK), under voucher numbers NMK:BCT:80001 to NMK:BCT:80221.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.489357/full#supplementary-material

Supplementary Datasheet 1—Config file for OTU analysis | The config file which was used to run an OTU analysis with *Cascabel* on the example 16S dataset.

Supplementary Datasheet 2—Config file for ASV analysis | The config file which was used to run an ASV analysis with *Cascabel* on the example 16S dataset.

Supplementary Datasheet 3—*Cascabel* library report | PDF report generated by *Cascabel* for the example 16S sequencing data analysis. The report contains the names and locations of all input and output files, names and short description of the modules ("rules") and parameters which were used in the analysis. In addition, graphics summarize the data in terms of sequence output and number of sequences left after each step of the analysis.

Supplementary Datasheet 4—*Cascabel* OTU report | PDF report describing the OTU analysis of the example 16S sequencing data. It contains the names and locations of all input and output files, names and short description of the modules ("rules") and parameters which were used in the analysis. In addition, graphics summarize the data in terms of sequence output per sample, number of OTUs and taxonomic composition.

Supplementary Datasheet 5—*Cascabel* ASV report | PDF report describing the ASV analysis of the example 16S sequencing data. It contains the names and locations of all input and output files, names and short description of the modules ("rules") and parameters which were used in the analysis. In addition, the data is summarized in terms of sequence output per sample, number of ASVs and taxonomic composition.

Supplementary Datasheet 6—Folder structure of results of an OTU analysis | This folder structure is generated for the output of an OTU analysis performed with *Cascabel*. Note that only the main files are displayed, log files and temporary files are not shown.

Supplementary Datasheet 7—Folder structure of results of an ASV analysis | This folder structure is generated for the output of an ASV analysis performed with *Cascabel*. Note that only the main files are displayed, log files and temporary files are not shown.

Supplementary Datasheet 8—Mock community datasets and evaluation | List of mock community datasets used, including NCBI BioProject and SRA accession numbers. Overview of *Cascabel* runs performed on the mock community data and evaluation of the performance of the individual runs.

Supplementary Datasheet 9—SRA identifiers of PRJNA526242 | Table spreadsheet indicating the sample, experiment and run identifiers of BioProject PRJNA526242 to download the raw data used for the worked example from the NCBI Sequence Read Archive (SRA).

Supplementary Datasheet 10—Config files of mock community analyses | Config files of the *Cascabel* runs performed on the mock community datasets as listed in **Supplementary Datasheet 8**.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amato, A., Kooistra, W. H. C. F., Ghiron, J. H. L., Mann, D. G., Pröschold, T., and Montresor, M. (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 158, 193–207. doi: 10.1016/j.protis.2006.10.001

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Anslan, S., Bahram, M., Hiiesalu, I., and Tedersoo, L. (2017). Pipecraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Mol. Ecol. Resour.* 17, e234–e240. doi: 10.1111/1755-0998.12692

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* 75, 129–137. doi: 10.3354/ame01753

Bakker, M. G. (2018). A fungal mock community control for amplicon sequencing experiments. *Mol. Ecol. Resour.* 18, 541–556. doi: 10.1111/1755-0998.12760

Balzano, S., Corre, E., Decelle, J., Sierra, R., Wincker, P., Da Silva, C., et al. (2015). Transcriptome analyses to investigate symbiotic relationships between marine protists. *Front. Microbiol.* 6:98. doi: 10.3389/fmicb.2015.00098

Balzano, S., Lattaud, J., Villanueva, L., Rampen, S. W., Brussaard, C. P., Bleijswijk, J., et al. (2018). A quest for the biological sources of long chain alkyl diols in the western tropical North Atlantic Ocean. *Biogeosciences* 15, 5951–5968. doi: 10.5194/bg-15-5951-2018

Balzano, S., Percopo, I., Siano, R., Gourvil, P., Chanoine, M., Marie, D., et al. (2017). Morphological and genetic diversity of beaufort sea diatoms with high contributions from the chaetoceros neogracilis species complex. *J. Phycol.* 53, 161–187. doi: 10.1111/jpy.12489

Besseling, M. A., Hopmans, E. C., Koenen, M., van der Meer, M. T., Vreugdenhil, S., Schouten, S., et al. (2019). Depth-related differences in archaeal populations impact the isoprenoid tetraether lipid composition of the Mediterranean Sea water column. *Organ. Geochem.* 135, 16–31. doi: 10.1016/j.orggeochem.2019.06.008

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

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., and et al. (2018). QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints* 6:e27295v2. doi: 10.7287/peerj.preprints.27295v2

Bradley, I. M., Pinto, A. J., and Guest, J. S. (2016). Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl. Environ. Microbiol.* 82, 5878–5891. doi: 10.1128/AEM.01630-16

Buza, T. M., Tonui, T., Stomeo, F., Tiambo, C., Katani, R., Schilling, M., et al. (2019). iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics* 20:374. doi: 10.1186/s12859-019-2965-4

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636

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

Caruso, V., Song, X., Asquith, M., and Karstens, L. (2019). Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems* 4:e00163-18. doi: 10.1128/mSystems.00163-18

Dong, X., Kleiner, M., Sharp, C. E., Thorson, E., Li, C., Liu, D., et al. (2017). Fast and simple analysis of miseq amplicon sequencing data with MetaAmp. *Front. Microbiol.* 8:1461. doi: 10.3389/fmicb.2017.01461

Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J., and Cordier, T. (2019). SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics* 20:88. doi: 10.1186/s12859-019-2663-2

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

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

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Forster, D., Dunthorn, M., Mahé, F., Dolan, J. R., Audic, S., Bass, D., et al. (2016). Benthic protists: the under-charted majority. *FEMS Microbiol. Ecol.* 92:fiw120. doi: 10.1093/femsec/fiw120

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protoc.* 2010:pdb-prot5368. doi: 10.1101/pdb.prot5368

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* 34, 942–949. doi: 10.1038/nbt.3601

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013). The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41, D597–D604. doi: 10.1093/nar/gks1160

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., et al. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Mol. Ecol.* 25, 3101–3119. doi: 10.1111/mec.13660

Hoppenrath, M., Beszteri, B., Drebes, G., Halliger, H., Van Beusekom, J. E. E., Janisch, S., et al. (2007). Thalassiosira species (Bacillariophyceae, Thalassiosirales) in the North Sea at Helgoland (German bight) and sylt (North

Frisian Wadden Sea)–a first approach to assessing diversity. *Eur. J. Phycol.* 42, 271–288. doi: 10.1080/09670260701352288

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jeraldo, P., Kalari, K., Chen, X., Bhavsar, J., and Mangalam, A. (2014). IM-TORNADO: a tool for comparison of 16S reads from paired-end libraries. *PLoS ONE* 9:e114804. doi: 10.1371/journal.pone.0114804

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10:5029. doi: 10.1038/s41467-019-13036-1

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Klunder, L., Duineveld, G. C., Lavaleye, M. S., van der Veer, H. W., Palsbøll, P. J., and van Bleijswijk, J. D. (2019). Diversity of Wadden Sea macrofauna and meiofauna communities highest in DNA from extractions preceded by cell lysis. *J. Sea Res.* 152:101764. doi: 10.1016/j.seares.2019.101764

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., et al. (2016). Open-source sequence clustering methods improve the state of the art. *mSystems* 1:e00003-15. doi: 10.1128/mSystems.00003-15

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Köster, J., and Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

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

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., et al. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* 17, 4035–4049. doi: 10.1111/1462-2920.12955

Morais, D., Roesch, L. F. W., Redmile-Gordon, M., Santos, F. G., Baldrian, P., Andreote, F. D., et al. (2018). BTW-bioinformatics through windows: an easy-to-install package to analyze marker gene data. *PeerJ* 6:e5299. doi: 10.7717/peerj.5299

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Needham, D., and Fuhrman, J. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat. Microbiol.* 1:16005. doi: 10.1038/nmicrobiol.2016.5

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

Parada, A. E., Needham, D. M., and Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* 18, 1403–1414. doi: 10.1111/1462-2920.13023

Parikh, H. I., Koparde, V. N., Bradley, S. P., Buck, G. A., and Sheth, N. U. (2016). MeFiT: merging and filtering tool for illumina paired-end reads for 16S rRNA amplicon sequencing. *BMC Bioinformatics* 17:491. doi: 10.1186/s12859-016-1358-1

Poncheewin, W., Hermes, G. D. A., van Dam, J. C. J., Koehorst, J. J., Smidt, H., and Schaap, P. J. (2019). NG-Tax 2.0: a semantic framework for high-throughput amplicon analysis. *Front. Genet.* 10:1366. doi: 10.3389/fgene.2019.01366

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Pylro, V. S., Morais, D. K., de Oliveira, F. S., Dos Santos, F. G., Lemos, L. N., Oliveira, G., et al. (2016). BMPOS: a flexible and user-friendly tool sets for microbiome studies. *Microb. Ecol.* 72, 443–447. doi: 10.1007/s00248-016-0785-x

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

Rimet, F., Kermarrec, L., Bouchez, A., Hoffmann, L., Ector, L., and Medlin, L. K. (2011). Molecular phylogeny of the family Bacillariaceae based on 18S rDNA sequences: focus on freshwater Nitzschia of the section Lanceolatae. *Diatom Res.* 26, 273–291. doi: 10.1080/0269249X.2011.597988

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

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

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

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

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849. doi: 10.1099/00207713-44-4-846

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

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

van Bleijswijk, J. D., Engelmann, J. C., Klunder, L., Witte, H. J., Witte, J. I., and van der Veer, H. W. (2020). Analysis of a coastal North Sea fish community: comparison of aquatic environmental DNA concentrations to fish catches. *Environ. DNA.* 2, 429–445. doi: 10.1002/edn3.67

van Bleijswijk, J. D. L., Whalen, C., Duineveld, G. C. A., Lavaleye, M. S. S., Witte, H. J., and Mienis, F. (2015). Microbial assemblages on a cold-water coral mound at the SE Rockall Bank (NE Atlantic): interactions with hydrography and topography. *Biogeosciences* 12, 4483–4496. doi: 10.5194/bg-12-4483-2015

van Bree, L., Peterse, F., Van der Meer, M., Middelburg, J., Negash, A., De Crop, W., et al. (2018). Seasonal variability in the abundance and stable carbon-isotopic composition of lipid biomarkers in suspended particulate matter from a stratified equatorial lake (Lake Chala, Kenya/Tanzania): implications for the sedimentary record. *Q. Sci. Rev.* 192, 208–224. doi: 10.1016/j.quascirev.2018.05.023

Vetrovský, T., Baldrian, P., and Morais, D. (2018). SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. *Bioinformatics* 34, 2292–2294. doi: 10.1093/bioinformatics/bty071

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Westcott, S. L., and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487. doi: 10.7717/peerj.1487

www.nature.com (2019). *Reporting Standards and Availability of Data, Materials, Code and Protocols*. Available online at: https://www.nature.com/nature-research/editorial-policies/reporting-standards (accessed July 12, 2019).

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., et al. (2020). PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *Gigascience* 9:giaa022. doi: 10.1093/gigascience/giaa022

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership