AN ABSTRACT OF THE THESIS OF

Ian R. Humphreys for the degree of Master of Science in Microbiology presented on May 29, 2019.

Title: Characterizing the Accuracy of Phylogenetic Analyses that Leverage 16S rRNA Sequencing Data.

Abstract approved:

_____

Thomas J. Sharpton

Investigations of 16S rRNA gene sequences hallmark modern microbiology. These sequences provide culture-independent insight into the abundance and distribution of microbiota and serve as a principle resource through which microbial community diversity is measured. Consequently, researchers rely on 16S gene sequences to test hypotheses rooted in ecology, evolution, and disease. Within 16S gene analyses, there exist potential sources of error that are often overlooked and under considered when developing studies and interpreting data. Prior research demonstrates that methodological sources of error introduced into 16S gene studies may arise from choices in sample preservation and storage temperature, DNA extraction method, PCR, and sequencing platform. Further variation can be introduced during informatic processing that is applied post DNA sequencing. Collectively, these errors limit the power of inferences derived from 16S rRNA gene sequences. It is therefore imperative to understand how study methodology impacts nucleotide sequence data to accurately interpret results from 16S genes. I provide a

summary of these methodological sources of error from literature and distill out best practices for conducting 16S rRNA studies when applicable. One widespread application of 16S rRNA sequences that microbiome studies frequently rely on is phylogenetic measures, which can assess microbial community diversity or infer evolutionary patterns. The conclusions drawn from these phylogenetic metrics assume the underlying phylogeny is reconstructed accurately; yet, the accuracy of phylogenetic trees has been shown to be dependent on a myriad of conditions, some of which remain unresolved. I describe how sequence length, region of the 16S gene, sequence diversity, and sample size effect the accuracy of 16S rRNA gene phylogenies using simulated data. Additionally, I show how incorporating full-length sequences selected from referential 16S rRNA sequence databases during phylogenetic reconstruction can improve the accuracy of 16S rRNA gene trees that are otherwise assembled from the short DNA sequences obtained by contemporary sequencing platforms. Collectively, I highlight through literature review the importance of experimental design throughout the typical steps taken during the 16S rRNA gene sequencing workflow, and I demonstrate through simulation analyses how several of these methodological choices impact the accuracy of resulting phylogenies.

Characterizing the Accuracy of Phylogenetic Analyses that Leverage 16S rRNA Sequencing Data

by
Ian R. Humphreys

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented May 29, 2019
Commencement June 2019

Master of Science thesis of Ian R. Humphreys presented on May 29, 2019

APPROVED:

Major Professor, representing Microbiology

Head of the Department of Microbiology

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Ian R. Humphreys, Author

ACKNOWLEDGEMENTS

First and foremost, I thank Dr. Thomas Sharpton for his unwavering support throughout my tenure in his lab. It is difficult to express in words how grateful I am for the mentorship he has provided that has enabled me to navigate through my first graduate school experience and independent research project. With his help, I feel prepared and excited to begin the next steps in what I hope to be a career in academia by pursuing a Ph.D.

I wish to thank the members of the Sharpton lab for creating an environment that I have thoroughly enjoyed working in. I owe a special thanks to Dr. Christopher Gaulke for his technical assistance, generosity, and advice. I thank Courtney Armour for her assistance despite having to endure my chronic harassment. As the completion of this degree marks the end of my time at Oregon State University, I wish to thank Dr. Jane Ishmael and members of the Ishmael laboratory: Dr. Jeffery Serrill, Dr. Xuemei Wan, and Daphne Mattos for fostering my passion for research as an undergraduate. Additionally, I would like to thank Dr. Jeff Chang, Dr. Kimberly Halsey, and Dr. Rebecca Vega Thurber for their time, enthusiasm, and vital roles in my graduate training as members of my thesis committee.

Finally to my parents Jonathan and Karen Humphreys and family for their endless love and support that I know will follow me anywhere.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

A Review of 16S Methodological Sources of Error and Variation

CHAPTER ONE

Ian R. Humphreys

Introduction

Microorganisms have long been known to cause disease (Blevins and Bronze 2010), play vital roles in nutrient cycles (eg. Falkowski, 1997), and benefit plants (eg. Freiberg et al. 1997) and animals (eg. Round and Mazmanian 2009). However, until recently, enormous sums of microbial diversity have gone unseen, unidentified, and unstudied. The technological advances that occurred at the turn of the century which resulted in high-throughput nucleic acid sequencing have unlocked rapid access to these diverse microbial communities and with it a revolution in microbiology. Now, complex assemblages of microbes that would once have been difficult to tease apart can be characterized using genomic techniques. As a result, ambitious large-scale projects such as the Earth Microbiome Project (Thompson et al. 2017), Human Microbiome Project (Huttenhower et al. 2012), American Gut Project (McDonald et al. 2018), and TARA Oceans (Sunagawa et al. 2015) have sought to study the microbiota associated with environmental and host-associated biomes. One key approach used to classify microbial taxa in these studies is the sequencing of universally conserved, taxonomically diagnostic "barcode" genes, such as that of the small subunit ribosomal RNA (16S rRNA gene in bacteria and archaea).

By simultaneously sequencing the 16S rRNA gene of the various taxa that comprise a microbial community, researchers can quickly and inexpensively determine which organisms comprise the community, quantify community biodiversity, and measure the phylogenetic relatedness of these organisms. The 16S rRNA gene encodes a critical component of the 30S small subunit of ribosomes, which is required by all known bacterial and archaeal cells due to the ribosome's essential role in translating mRNA into

protein. Consequently, the 16S rRNA gene serves as a wide-spread genetic marker that can be used to discriminate between microbial taxa. The structure of the 16S rRNA gene and conservation of function constrain the rate of mutations throughout the gene. This results in nine hypervariable regions flanked by regions of high conservation (Woese et al. 1980 and Yang et al. 2016). Due to the vastly different rates of evolution, polymorphisms in the hypervariable regions can be used to distinguish between more recently diverged lineages while those in the highly conserved regions can be used to infer more ancient divergences. PCR primers have been designed to target highly conserved regions to amplify adjoining hypervariable regions for DNA sequencing (Baker et al. 2003).

Degenerative PCR primers that target 16S rRNA gene hypervariable regions enable rapid and inexpensive insights into the taxa that comprise microbial communities. Studies that rely on 16S gene sequences typically include several steps. First, environmental or host-associated samples such as soil or feces that contain microorganisms are collected. The DNA from each sample is then extracted and PCR primers are often used to amplify 16S rRNA gene regions of interest prior to DNA sequencing. Finally, the resulting 16S gene sequences are used to test hypotheses.

Yet, despite the power genome-based studies offer, there exists potential for errors. Differences in methodology such as DNA extraction, amplification, and bioinformatic pipelines can result in artificial variation that is comparable to biological variation that can be measured between different samples (Sinha et al. 2017). Additionally, reproducibility between 16S rRNA gene studies has been shown to vary between facilities (HMP Consortium 2012; Sinha et al. 2016). These effects may obscure

biologically significant signal, result in false discoveries, and influence downstream analyses. Therefore, it is important to consider how study parameters impact 16S rRNA gene sequence-based investigations. In the following subsections, we review the state of knowledge about how various study parameters, including sample storage and preservation, PCR, primer selection, sequencing platform, and bioinformatics pipelines on inferences about bacterial communities.

Sample Storage

Environmental and host-associated samples that contain microbial biomass may be collected long before DNA extraction. Buffering solutions and temperature controlled storage stabilize the DNA of microbial communities to ensure accurate detection of microbial taxa at a later date. However, different preservation methods and storage temperatures can produce inherent biases in 16S rRNA gene-based studies.

An array of preservation techniques have been developed and several studies have assessed their impact on the accuracy of estimates of community composition and diversity. OMNIgene.GUT buffer resulted in lower compositional changes in gut microbiome communities compared to fresh samples than RNAlater, 70% ethanol, 95% ethanol, and Whatman FTA cards (Song et al. 2016). Another study showed that samples preserved in OMNIgene.GUT were more similar to cold-stored samples, which are generally considered to stabilize DNA, than replicates stored in RNAlater, Tris-EDTA, or at room-temperature (Choo et al. 2015). When cooling is unavailable, card-based preservation methods such as fetal occult blood test (FOBT) or Whatman FTA cards may be better choices than buffer solutions (Sinha et al. 2016; Dominianni et al. 2014; Song et

al. 2016). We caution against the use of RNAlater because its use may result in decreased DNA purity and lower microbial diversity (Dominianni et al. 2014), higher variation in microbial communities with heat (Song et al. 2016), and reduced DNA yields (Gorzelak et al. 2016). Similarly, preservation in 70% ethanol may decrease community stability with heat (Sinha et al. 2016; Song et al. 2016). As a result, if ethanol preservation is used, concentrations should be at least 95% ethanol and regardless of preservation method, samples should be stored at cold temperatures (Song et al. 2016; Hale et al. 2015).

Additional considerations when selecting between available preservation techniques include potential for conducting further analyses. For example samples stored in RNAlater can be used for transcriptomic investigations and samples stored in ethanol can be used for metabolomics studies (Sinha et al. 2016). Therefore sample preservation method should likely be considered based on the planned analyses that will be conducted to investigate a biological question of interest.

In conjunction with sample preservation methods, storing samples in temperature controlled environments can reduce variation in microbial communities that can occur over time. Generally, -80°C storage of biological and environmental samples is regarded as the highest fidelity storage temperature to preserve DNA quality and ensure accurate microbial community profiles (Choo et al. 2015; Lauber et al. 2010; Bahl et al. 2012; Tzeneva et al. 2009; Fouhy et al. 2015). Microbial communities from gut microbiome samples have low differences in β-diversity between storage temperatures but greater differences in abundance weighted β-diversity (Song et al. 2016). Relative abundance estimates have been shown to vary by sample storage temperature by multiple groups ranging from -80°C to approximately 25°C (Roesh et al. 2009; Choo et al. 2015; Lauber

et al. 2010; Bahl et al. 2012, Gorzelak et al. 2015). Of potential interest to gut microbiome researchers, two gut microbiome studies reported significant shifts in the abundance of Firmicutes and Bacteroidetes between samples stored in differing temperatures (Bahl et al. 2012; Gorzelak et al. 2015). Variation in the ratios of these phyla may obscure biologically meaningful results because the ratio of Bacteriodetes to Firmicutes in fecal samples is frequently treated as an indicator of host health (Ley et al. 2006; Koliada et al. 2017). Conversely, other studies have found that there are no significant differences between the relative abundance of major phyla in gut microbiome samples stored in differing temperatures without buffers or subjected to two thaw cycles (Dominianni et al. 2014; Bassis et al. 2017).

However, the effects of storage temperature on microbial communities may be environment specific. Minimal variation in microbial community composition are associated with storage temperatures for human oral microbiome samples (Lou et al. 2014), skin microbiome samples (Lauber et al. 2010), and vaginal microbiome samples (Bai et al. 2012) stored in buffer solutions. For environmental samples, the community composition of soil stored at room temperature for up to 14 days were mostly unaffected (Lauber et al. 2010); however, air-dried soil samples stored for three months exhibited significant differences in richness and diversity of bacterial profiles compared to samples stored at -80°C (Tzeneva et al. 2009).

In summary, based on the findings of our literature search, when samples cannot be processed shortly after collection, OMNIgene.GUT buffer solution and -80°C storage of microbial samples provide the most protection for microbial DNA and minimize shifts in community composition. That said, new preservation methods have recently entered

the market and may provide improved results. Some microbial communities or samples collected from certain environments may benefit less from low temperature storage than others due to naturally lower moisture content or increased variation in environmental conditions. Due to the diversity of DNA preservation methods employed in conjunction with temperature storage it is difficult to disentangle absolute guidelines. Further work should be conducted to elucidate the effects of sample preservation and long-term storage strategy on the integrity of microbial community DNA across different microbiomes. However regardless of methodology, we stress the importance of consistency across samples to reduce batch effects.

DNA Extraction

DNA yield is dependent on the method of storage, preservation, and extraction techniques (Nechvatal et al. 2008). DNA isolation can be conducted through the use of classical techniques such as phenol-chloroform or chaotropic salts based extractions. However in the age of high-throughput sequencing, biotech companies have engineered all-inclusive kits to expedite extractions and standardize methodology. Depending on the extraction method, researchers have reported varying yields of DNA and purity of nucleic acids which have been shown to result in differing community diversity and abundance estimates. Yet despite improvements, regardless of method, biases are introduced during DNA extraction (Yuan et al. 2012; Brooks et al. 2015) and must be considered in study design.

Choice of DNA extraction method affects the overall DNA concentration obtained from samples; however, studies conflict as to which method recovers the most

accurate and highest quality DNA. In human fecal samples, use of the QIAamp DNA Stool Kit (QIAGEN) for DNA extractions was shown to produce higher average DNA yields than extractions using the MoBio Fecal Kit (now owned by QIAGEN) (Nechvatal et al. 2008). Additionally, use of the QIAamp DNA Stool mini Kit for extracting DNA produces better nucleic acid purity, greater sequencing yield, longer reads after quality trimming, and higher OTU-level diversity than phenol-chloroform or chaotropic salt based DNA extractions yet lower double stranded DNA yield than chaotropic salt DNA extractions (Gerasimidis et al. 2016). Conversely, in other gut microbiome studies, use of the PowerSoil DNA Isolation Kit (now owned by QIAGEN as PowerFecal Kit) resulted in higher DNA yield than QIAamp DNA Stool Kit (Bahl et al. 2012) and outperformed the QIAamp DNA Stool Kit in low bacterial biomass samples (Velásquez-Mejia et al. 2018).

Estimates of relative abundance for microbial taxa may be biased by DNA extraction method (Brooks et al. 2015; Velasquez-Mejia et al. 2018; Yuan et al. 2012). For example, use of the MoBio PowerSoil DNA Isolation Kit resulted in an increased number of Firmicutes and Actinobacteria and a decrease in Bacteroidetes compared to samples extracted using a QIAamp DNA Stool mini Kit (Velasquez-Mejia et al. 2018). Depending on the physical properties of the microorganisms present in the sample, DNA extractions that incorporate standard chemical lysis may be unable to access DNA from the whole microbial community. Organisms such as *Mycobacterium spp.* and *Bacillus* can form spores which contain thick cell walls that require mechanical lysis techniques to recover DNA (Kuske et al. 1998; Vandeventer et al. 2011). As a result, mechanical lysis is considered a necessary component of DNA extraction that can be added to any DNA

extraction protocol through a bead beating preprocessing step. Bead beating has been shown to reduce biases during DNA extraction that effect downstream community calculations of richness and relative abundance estimations due to the inability to access DNA from subsets of bacterial and archaeal populations (Kuske et al. 1998; Carrigg et al. 2007; Yuan et al. 2012; Salonen et al. 2010; de Boer et al. 2009; Smith 2011). While there are a multitude of different options for mechanical lysis, 0.1 mm silica beads have been shown to improve the recovery of Gram positive bacteria during DNA extraction without negatively impacting Gram negative organisms (de Boer et al. 2009).

As a result, we recommend mechanical lysis if not already integrated into the kit protocol to maximize microbial diversity recovered from samples and minimize additional taxa specific biases during DNA extraction. It is difficult to identify a single optimal DNA extraction method however, use of standardized kit-based methods improve reproducibility. Additionally, we stress that a single method of DNA extraction should be executed for a given study to negate inter-sample biases.

Hypervariable Region Selection

The vast majority of 16S rRNA-based studies use polymerase chain reaction (PCR) to target and amplify specific regions of the 16S gene due to technological limitations of the most widely used sequencing platforms that result in short length sequences. Therefore until long-read sequencing that spans whole genes is universally adopted, primers will continue to be used to PCR amplify regions of the 16S gene that possess high nucleotide variation which allow for differentiating between taxa. As a result, short regions of the 16S gene are used to approximate the variation encompassed

in the roughly 1500 nucleotide long gene that itself only represents a small portion of an organism's genome. To target hypervariable regions, primers bind to complementary highly conserved sequences in one of the highly conserved regions of the 16S rRNA gene that flank each hypervariable region (Baker et al. 2003). Rates of nucleotide conservation and hypervariable region length vary which consequently dictates the efficacy of each region to differentiate between taxa. Researchers have extensively considered how the use of DNA sequences from the different hypervariable regions impact study outcomes such as phylogeny-based measurements, taxonomic classification rates, and community diversity metrics. Phylogenies reconstructed using V4-V6 region sequences which encompass hypervariable regions four through six (Yang et al. 2016) and V3/V4 sequences (Ragan-Kelley et al. 2013) are most representative of full-length 16S phylogenies while V2 and V8 (Yang et al. 2016) and V9 (Ragan-Kelley et al. 2013) were least similar to the full-length phylogenies. For taxonomic classification, V4 hypervariable region sequences are, on average, best able to assign sequences genus level taxonomic labels across different sampling environments (Soergel et al. 2012). β-diversity metrics applied to 16S data were robust to primer and sequencing platform selection; however, primer choice influences relative abundance estimations (Tremblay et al. 2015). Of those tested (V4, V6-V8, and V7-V8), the V4 hypervariable region sequences most closely resemble community profiles obtained using shotgun sequencing (Tremblay et al. 2015). Simulated V4, V5-V6, and V6-V7 hypervariable region fragments most closely estimate the full-length 16S sequence species richness (Youssef et al. 2009). Despite numerous studies identifying the "best" hypervariable region(s) that most closely resemble results obtained from using full-length 16S gene sequences, it is

difficult to determine which primer pair to use because of PCR biases that affect taxa unequally.

Polymerase Chain Reaction

PCR is a widely used step in 16S rRNA gene sequencing which enables the analysis of low biomass samples by amplifying specific segments of DNA. Unfortunately, errors can occur during PCR, and these errors can compound with each additional amplification cycle. For example, poor DNA polymerase fidelity can result in substitutions, insertions, and deletions as well as off-target primer binding, which may result in chimeras that arise from incompletely extended sequences annealing to another sequence. These PCR errors can significantly impact estimation of community diversity and composition.

Use of high fidelity DNA polymerases such as KAPA and minimizing the number of PCR rounds help to alleviate the formation of chimeras, nucleotide polymorphisms, and compositional biases in microbial communities (Gohl et al. 2016; Sze and Schloss 2019). Sze and Schloss used mock communities to demonstrate that the number of PCR rounds is of primary importance and polymerase choice is secondary (Sze and Schloss 2019). After clustering sequences to reduce noise, at 30 rounds of PCR amplification, KAPA polymerase had the lowest error rate followed by Phusion, Q5, Accuprime, and Platinum; however, Accuprime had the fewest chimeras followed by KAPA, Phusion, Q5 and Platinum (Sze and Schloss 2019). As additional rounds of PCR are conducted, Shannon diversity index generally increased and bacterial communities became more even (Sze and Schloss 2019). As a result, Sze and Schloss caution against comparing data

from differing PCR conditions (Sze and Schloss 2019). Gohl and colleagues found that beyond 20 cycles of PCR, KAPA polymerase out performs Q5 and Taq both in having the lowest nucleotide error rate and least number of chimeric sequences (Gohl et al. 2016). Additionally, reducing the amount of starting DNA used in PCR decreases the percentage of chimeric reads detected after DNA sequencing (D'Amore et al. 2016).

Sequencing Technology

Long gone are the days of determining DNA sequences from 2D gel electrophoresis, Sanger sequencing, and most recently Roche 454. Instead, Illumina's HiSeq and MiSeq sequencing platforms have quickly become the sequencing standard as they have been shown to produce higher quantity and quality reads than Roche 454 (Caporaso et al. 2012). Yet researchers must still select an appropriate sequencing platform and understand the benefits and weaknesses associated with their decision.

The two Illumina sequencers: HiSeq and MiSeq can be distinguished from each other by scale of operation, cost, and read length. MiSeq machines deliver rapid smaller scale sequencing while the HiSeq reduces per sample cost by enabling higher parallelization at the expense of time and sequence length (Caporaso et al 2012). MiSeq and HiSeq have both been shown to produce low variability across lanes in a single run and similar quality reads (Caporaso et al. 2012). Taking advantage of the higher quantity of reads, dual-index paired-end primers have enabled MiSeq reads to attain similar error rates to Roche 454 GS-FLX Titanium while increasing read-depth by 10-fold (Kozich et al. 2013). Unfortunately, MiSeq is currently limited to short read sequencing of roughly

300 nucleotides and attempts to increase read length generally result in reduced overlap between reads that limit the ability to correct errors (Schloss et al. 2016).

Illumina's HiSeq and MiSeq platforms are limited to short sequence lengths which has forced investigators to focus on the short information rich hypervariable regions of the 16S gene. Emerging long-read sequencing technologies such as PacBio and Oxford Nanopore hold potential to transform 16S investigations by offering access to full-length gene sequence reads. For example, longer sequences are more likely to receive better resolved taxonomic annotations to the level of genus or species (Schloss et al. 2016 and Pootakham et al. 2017) and reconstruct phylogenies more similar to those reconstructed using full-length genes (Ragan-Kelley et al. 2013). One limitation of long-read sequencing technologies that has reduced their adoption is concern surrounding their higher sequencing error rates compared to the HiSeq and MiSeq. That said, these technologies are rapidly improving and new informatic solutions targeted at reducing long-read errors are being developed. For example, after conducting read filtering and quality control, PacBio (P6-C4 chemistry) can produce sequences with error rates of around 0.03% (Schloss et al. 2016 and Wagner et al. 2016). Another potential effect of long-read sequencing is on the accuracy of estimates of species richness. One study found that MiSeq V1-V2 sequences have elevated species richness estimates than PacBio full-length sequences from the same sample (Wagner et al. 2016). However, when the full-length PacBio sequences were truncated to simulate V1-V2 reads, there was an increase in species diversity indicating that short read sequencing may result in overestimation of species diversity (Wagner et al. 2016).

As new sequencing platforms are developed and chemistries improve, the per nucleotide error rates resulting from sequencing error will likely decrease. Currently, a large factor in platform selection resides in cost, wherein HiSeq is the cheapest followed by MiSeq and then PacBio. Unfortunately, read length and read quality are proportional to cost. Therefore, the selection of a sequencing platform should be based on experimental need. The following sections which discuss downstream bioinformatic analyses may provide additional insight into which sequencing platform should be utilized choice.

Bioinformatics

DNA sequencers produce "raw" reads which must be subject to computational quality control prior to analysis. During this bioinformatic cleanup process, there exist numerous options in software each designed to produce optimal results for differing scenarios. This section provides an overview of important steps in 16S gene sequence processing pipelines and highlights examples of stand-alone and popular all-inclusive methods.

First, sequencing adaptors must be removed from raw amplicon reads prior to their subsequent analysis (e.g., cutadapt (Martin 2011)). Reads are then typically subject to paired-end assembly (e.g., PANDAseq (Masella 2012)), which merges mate pairs into longer 16S rRNA gene contigs, as well as quality trimming (e.g., Cutadapt (Martin 2011) or Sickle (Joshi and Fass 2011)), which filters or truncates error prone read sequences. Chimeras are then identified and removed from the set of reads (e.g., UCHIME (Edgar et al. 2011) or DECIPHER (Wright et al. 2012)).

After these quality filtering steps, sequences can be assigned into operational taxonomic units (OTUs) in three general ways; *de novo*, reference-based, and open-reference. Although OTUs can be created in different ways, studies have demonstrated that *de novo* methods which do not rely on information from a database outperform reference-based clustering that leverage database-dependent taxonomy binning (Schloss and Westcott 2011; Westcott and Schloss 2015; Schloss 2016). Furthermore, between different *de novo* based methods, average neighbor clustering which averages the differences between pairs of sequences was the most robust method (Schloss and Westcott 2011; Schloss 2016). Additionally, when OTU clustering was applied to human twin gut microbiomes, *de novo* clustering identified a higher number of heritable OTUs between twin pairs than other approaches (Jackson et al. 2016) which improved the power of the analysis. Amplicon sequence variants (ASVs) generated by DADA2 provide an alternative sequence clustering method that applies sequencing run-specific error model training to reduce sequencing-error and preserve fine-scale variation between sequences that may be lost during OTU clustering (Callahan et al. 2016).

ASV or OTU-clustered representative sequences are then aligned to enable comparisons between the sequences, assign taxonomy, or construct phylogenetic trees. Three primary algorithms that are commonly used in nucleotide alignments: *de novo* pairwise, *de novo* multiple sequence, and profile-based alignments each offer differing levels of speed and accuracy (Schloss 2009). Before or after alignment, sequences can be taxonomically annotated using SILVA (Yarza et al. 2008), Greengenes (DeSantis et al. 2006), or RDP (Cole et al. 2009) 16S rRNA sequence databases. Each 16S database contains sequences with varying levels of alignment quality and phylogenetic diversity

(Schloss 2010) that result in environment specific taxonomic classification accuracy. For example, SILVA-based taxonomic classification classifies human fecal microbiomes and soil samples with greater accuracy than Greengenes or RDP while RDP-based taxonomic classification better classifies mouse feces (Schloss et al. 2016).

Rather than create custom software pipelines to string together these vital informatic processes, there exist several software packages that expedite these steps and bring added uniformity between studies. Of the most commonly used software suits, mothur (Schloss et al. 2009) and QIIME (Caporaso et al. 2010) are OTU-based while DADA2 (Callahan et al. 2016) and most recently QIIME 2 (Bolyen et al. 2018) produce ASVs. In the end, regardless of the sequencing technology and software selection, inclusion of quality trimming, error correction and read assembly can significantly reduce substitution errors (Schirmer et al. 2015).

Phylogenetics

Once sequences are processed, filtered, clustered into ASVs or OTUs, and accurately aligned, phylogenies can be reconstructed to provide additional insights into microbial communities. Phylogenetic trees allow for the calculation of evolutionarily informed measures of β-diversity (Lozupone and Knight 2005), identification of phylogenetic and co-phylogenetic signal (Gaulke et al. 2018), and trait identification (Washburne et al. 2017). Yet phylogenetic trees have been shown to vary based on gene, region, sequence length, alignment, diversity, and reconstruction method. To draw meaningful conclusions from these tools which rely on phylogenies, researchers must be aware of the methodological sources of phylogenetic error that may impact their results.

Depending on the study, differing gene segments may result in improved taxonomic resolution. For example, 16S rRNA is known to be unable to differentiate between Bacteroidaceae and Bifidobacteriaceae (Moeller et al. 2016) thus alternative markers should be used when taxa of biological interest are known to have poor separation with 16S gene sequences. Longer sequences are better able to recapitulate full-length genetic variation (Schloss 2010), increase the proportion of correct trees (Graybeal 1998), improve branch-length calculations (Rosenberg and Kumar 2003), and more accurately represent the phylogenetic distance of full-length phylogenies (Ragan-Kelley et al. 2013). However, due to potentially uninformative stretches within genes, analyzing the appropriate region(s) of a gene that yield discriminatory power between taxa has a greater effect on phylogenetic inferences than increasing sequence length (Martin et al. 1995). Despite longer sequences improving results, it is critical to trim sequences to the same starting and ending regions to ensure phylogenetic accuracy because different regions of genes do not mutate at uniform rates (Schloss 2010). The ability of different 16S hypervariable regions to compute community diversity metrics is discussed in a prior section.

There are four primary types of phylogenetic reconstruction methods that model evolutionary relationships from aligned sequences: distance, parsimony, and maximum likelihood and Bayesian inference. Distance based methods such as neighbor joining (Saitou and Nei 1987) or minimum-evolution (Rzhetsky and Nei 1992) rely on a distance matrix composed of all taxa. Maximum parsimony methods minimize the number of evolutionary events predicted in the final phylogeny (Felsenstein 2004). Both maximum likelihood and Bayesian inference employ probability-based statistical approaches to

determine the optimal tree. Maximum likelihood methods determine the tree that has the highest probability of depicting evolutionary history based on the likelihood function while Bayesian inference uses posterior probabilities to optimize topology (Svennblad et al. 2006).

The accuracy of reconstruction method depends on substitution rate, number of sites, and number of taxa (Rosenberg and Kumar 2001; 2003). Generally, maximum likelihood and Bayesian methods reconstruct phylogenies most accurately followed by maximum parsimony and neighbor-joining (Rosenberg and Kumar 2001; Ogden and Rosenberg 2006; Price et al. 2010). Currently, some of the most popular software used in microbiome studies for phylogenetic tree reconstruction are FastTree2 (Price et al. 2010), RaxML (Stamatakis 2012), and BEAST (Drummond and Rambaut 2007). Recently released RaxML-NG appears promising as it boasts the improved accuracy of maximum likelihood with greatly reduced computational time compared to prior options (Kozlov et al. 2019).

While different methods of phylogenetic tree reconstruction will provide varying levels of accuracy, phylogenies in general are highly dependent on the quality of sequence alignment. Morrison and Ellis found that sequence alignments accounted for more phylogenetic variation than tree-building method (Morrison and Ellis 1997). Schloss has conducted extensive studies that demonstrate differences in alignment quality between full-length 16S databases that are commonly used for reference-based alignment and found that poor quality alignments inflate phylogenetic diversity (Schloss 2009; 2010). As a result, the poor variable region alignments in Greengenes predict higher genetic diversity, richness, and phylogenetic diversity than SILVA and RDP alignments

(Schloss 2010). Furthermore, errors in topology from poor alignments become magnified in phylogenies with shallow diversity (Ogden and Rosenberg 2006). Additionally, sequence diversity and the number of lineages impact phylogenetic accuracy (reviewed in Hillis et al. 2003; Nabhan and Sarker 2012).

Overall, maximizing the accuracy of phylogenetic analyses is complex and requires researchers to understand how each decision in their analyses may affect potential conclusions. Generally, to improve phylogenetic accuracy the most important considerations are the gene region of interest and alignment algorithm. Secondarily, tree reconstruction method, sequence length, number of lineages, and diversity between lineages influence phylogenetic accuracy, Additional considerations must be made if conducting clade-based analyses due to their dependence on rooted phylogenies.

Conclusions

16S rRNA analyses provide powerful, inexpensive insights into microbial communities that may otherwise remain unexplored. Consequently, it is important for researchers to understand how sources of error such as PCR amplification bias and sequence error can accumulate in pyrosequencing studies and imperative to understand how these errors may be controlled for. Throughout our literature search we were unable to identify a universal consensus about the best methodological practices and community specific biases appear pervasive to microbiome studies. Despite this fact, there are several broad recommendations that can be relayed. Firstly, DNA extraction of fresh samples circumvents potential storage and preservation effects although in the event of delayed sample processing, cold storage reduces potential changes to the microbial community

composition. Secondly, kit-based DNA extraction methods can reduce variability and improve cross-study comparisons and mechanical lysis should be integrated to ensure maximum diversity within the community is captured. Thirdly, optimal primer selection may be microbial community specific, however, generally reads that include portions of the V4 hypervariable region appear to provide improved discriminatory power. Fourthly, during bioinformatic processing, we suggest careful attention during sequence alignment and appropriate selection of clustering dependent on the biological question of interest. Another consideration that has been posited to reduce methodological errors is to incorporate mock communities into sequencing studies. Mock communities can serve as a strong quality control to identify error-driven outliers within samples (Bender et al. 2018). Additionally, for meta-analyses, researchers must be cognizant of study-effects that may diminish cross-study comparisons. Finally, we stress that methodological consistency between samples within a study are of paramount importance to ensure that there are no methodological sample specific effects.

While prior work affords meaningful methodological recommendations for most of the steps associated with the generation and analysis of 16S rRNA gene sequence data, there remains many open questions that need to be answered to help ensure that future research is as accurate as possible. In the following chapter, I delve deeply into one specific underexplored area of bioinformatic analysis of 16s rRNA gene sequence data: phylogenetics. Surprisingly few of the phylogenetic analyses discussed prior were conducted on microbial genes and fewer yet specifically focused on the accuracy of 16S rRNA gene phylogenies. Therefore we sought to determine how phylogenetic accuracy is influenced by study design within the context of 16S rRNA microbial microbiomes. In

particular the following work seeks to address how sequence length, sample diversity, sample size, and hypervariable regions affect the accuracy of 16S phylogenetic trees. In doing so, we offer researchers insights into potential phylogenetic inaccuracies and hope these data will be taken into consideration during the design of study methods.

References

1. Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555.
2. Bai G, Gajer P, Nandy M, Ma B, Yang H, Sakamoto J, Blanchard MH, Ravel J, Brotman RM. 2012. Comparison of storage conditions for human vaginal microbiome studies. PLoS One 7:e36934.
3. Bahl MI, Bergström A, Licht TR. 2012. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. FEMS Microbiol Lett 329:193–197.
4. Bassis CM, for the CDC Prevention Epicenters Program, Moore NM, Lolans K, Seekatz AM, Weinstein RA, Young VB, Hayden MK. 2017. Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. BMC Microbiology.
5. Bender JM, Li F, Adisetiyo H, Lee D, Zabih S, Hung L, Wilkinson TA, Pannaraj PS, She RC, Bard JD, Tobin NH, Aldrovandi GM. 2018. Quantification of variation and the impact of biomass in targeted 16S rRNA gene sequencing studies. Microbiome.
6. Blevins SM, Bronze MS. 2010. Robert Koch and the "golden age" of bacteriology. International Journal of Infectious Diseases.
7. Bolyen E, Rideout JR, Dillon MR, Bokulich NA … Knight R, Caporaso JG. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints. 6:e27295v2.
8. Brooks JP, Paul Brooks J, Vaginal Microbiome Consortium (additional members), Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiology.
9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods.
10. Carrigg C, Rice O, Kavanagh S, Collins G, O'Flaherty V. 2007. DNA extraction method affects microbial community profiles from soils and sediment. Applied Microbiology and Biotechnology.
11. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods.
12. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal.
13. Choo JM, Leong LEX, Rogers GB. 2015. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:16350.

14. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research.
15. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics.
16. de Boer R, Peters R, Gierveld S, Schuurman T, Kooistra-Simid M, Savelkoul P. 2010. Improved detection of microbial DNA after bead-beating before DNA isolation. Journal of Microbial Methods. 80:2
17. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072.
18. Dominianni C, Wu J, Hayes RB, Ahn J. 2014. Comparison of methods for fecal microbiome biospecimen collection. BMC Microbiol 14:103.
19. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214
20. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics.
21. Falkowski PG. 1997. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. Nature.
22. Felsenstein J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.
23. Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Paul Ross R, O'Callaghan G, Plant BJ, Stanton C. 2015. The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. PLOS ONE.
24. Freiberg C, Fellay R, Bairoch A, Broughton WJ, Rosenthal A, Perret X. 1997. Molecular basis of symbiosis between Rhizobium and legumes. Nature 387:394–401.
25. Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. 2018. Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota. mBio.
26. Gerasimidis K, Bertz M, Quince C, Brunner K, Bruce A, Combet E, Calus S, Loman N, Ijaz UZ. 2016. The effect of DNA extraction methodology on gut microbiota research applications. BMC Research Notes.
27. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D, Beckman KB. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology.
28. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. 2015. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. PLoS One 10:e0134802.
29. Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47:9–17.

30. Hale VL, Tan CL, Knight R, Amato KR. 2015. Effect of preservation method on spider monkey (Ateles geoffroyi) fecal microbiota over 8weeks. Journal of Microbiological Methods.
31. Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? Systematic Biology.
32. The Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature. 486:215–221.
33. Huttenhower C … The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature.
34. Jackson MA, Bell JT, Spector TD, Steves CJ. 2016. A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. PeerJ.
35. Joshi NA and Fass JN. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.
36. Koliada A, Syzenko G, Moseiko V, Budovska L, Puchkov K, Perederiy V, Gavalko Y, Dorofeyev A, Romanenko M, Tkach S, Sineok L, Lushchak O, Vaiserman A. 2017. Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiology.
37. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology.
38. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics.
39. Kuske CR, Banton KL, Adorada DL, Stark PC, Hill KK, Jackson PJ. 1998. Small-Scale DNA Sample Preparation Method for Field PCR Detection of Microbial Cells and Spores in Soil. Appl Environ Microbiol 64:2463–2472.
40. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiology Letters.
41. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006. Human gut microbes associated with obesity. Nature.
42. Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology.
43. Luo T, Srinivasan U, Ramadugu K, Shedden KA, Neiswanger K, Trumble E, Li JJ, McNeil DW, Crout RJ, Weyant RJ, Marazita ML, Foxman B. 2016. Effects of Specimen Collection Methodologies and Storage Conditions on the Short-Term Stability of Oral Microbiome Taxonomy. Applied and Environmental Microbiology.
44. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL,

Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, Tripathi A, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, American Gut Consortium, Knight R. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 3.

45. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal.
46. Martin MJ, González-Candelas F, Sobrino F, Dopazo J. 1995. A method for determining the position and size of optimal sequence regions for phylogenetic analysis. J Mol Evol 41:1128–1138.
47. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 13:31.
48. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids. Science.
49. Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol 14:428–441.
50. Nabhan AR, Sarkar IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics.
51. Nechvatal JM, Ram JL, Basson MD, Namprachan P, Niec SR, Badsha KZ, Matherly LH, Majumdar APN, Kato I. 2008. Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. Journal of Microbiological Methods.
52. Ogden TH, Heath Ogden T, Rosenberg MS. 2006. Multiple Sequence Alignment Accuracy and Phylogenetic Inference. Systematic Biology.
53. Pootakham W, Mhuantong W, Yoocha T, Putchim L, Sonthirod C, Naktang C, Thongtham N, Tangphatsornruang S. 2017. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. Scientific Reports.
54. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE.
55. Ragan-Kelley B, Walters WA, McDonald D, Riley J, Granger BE, Gonzalez A, Knight R, Perez F, Gregory Caporaso J. 2013. Collaborative cloud-enabled tools allow rapid, reproducible biological insights. The ISME Journal.
56. Roesch LFW, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW. 2009. Influence of Fecal Sample Storage on Bacterial Community Diversity. The Open Microbiology Journal.
57. Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proceedings of the National Academy of Sciences.

58. Rosenberg MS, Kumar S. 2003. Taxon Sampling, Bioinformatics, and Phylogenomics. Systematic Biology.
59. Round JL, Mazmanian SK. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nature Reviews Immunology.
60. Rzhetsky A, Nei M. 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. Journal of Molecular Evolution.
61. Saitou N, Nei M. 1987. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol Biol. Evol.* 4(4):406-425.
62. Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajilić-Stojanović M, Kekkonen RA, Palva A, de Vos WM. 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. Journal of Microbiological Methods.
63. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research.
64. Schloss PD. 2009. A high-throughput DNA sequence aligner for microbial ecology studies. PLoS One 4:e8230.
65. Schloss PD. 2010. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLoS Computational Biology.
66. Schloss PD. 2016. Application of a Database-Independent Approach To Assess the Quality of Operational Taxonomic Unit Picking Methods. mSystems 1:2.
67. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ.
68. Schloss PD, Westcott SL. 2011. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. Applied and Environmental Microbiology.
69. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology.
70. Smith B. 2011. Optimising Bacterial DNA Extraction from Faecal Samples: Comparison of Three Methods. The Open Microbiology Journal.
71. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Microbiome Quality Control Project Consortium, Abnet CC, Knight R, White O, Huttenhower C. 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 35:1077–1086.

72. Sinha R, Chen J, Amir A, Vogtmann E, Shi J, Inman KS, Flores R, Sampson J, Knight R, Chia N. 2016. Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. Cancer Epidemiology Biomarkers & Prevention.
73. Soergel DAW, Dey N, Knight R, Brenner SE. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. The ISME Journal.
74. Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. 2016. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems.
75. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics.
76. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Boss E, Bowler C, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D, Coordinators TO. 2015. Structure and function of the global ocean microbiome. Science.
77. Svennblad B, Erixon P, Oxelman B, Britton T. 2006. Fundamental Differences Between the Methods of Maximum Likelihood and Maximum Posterior Probability in Phylogenetics. Systematic Biology.
78. Sze MA and Schloss PD. 2019. The Impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. Biorxiv.
79. Thompson LR, The Earth Microbiome Project Consortium, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature.
80. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771.
81. Tzeneva VA, Salles JF, Naumova N, de Vos WM, Kuikman PJ, Dolfing J, Smidt H. 2009. Effect of soil sample preservation, compared to the effect of other environmental variables, on bacterial and eukaryotic diversity. Res Microbiol 160:89–98.
82. Vandeventer PE, Weigel KM, Salazar J, Erwin B, Irvine B, Doebler R, Nadim A, Cangelosi GA, Niemz A. 2011. Mechanical Disruption of Lysis-Resistant

Bacterial Cells by Use of a Miniature, Low-Power, Disposable Device. Journal of Clinical Microbiology.
83. Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Escobar JS. 2018. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. Applied Microbiology and Biotechnology.
84. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiology.
85. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ.
86. Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3:e1487.
87. Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius R, Gutell R, Hogan JJ, Noller HF. 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Research.
88. Wright ES, Safak Yilmaz L, Noguera DR. 2012. DECIPHER, a Search-Based Approach to Chimera Identification for 16S rRNA Sequences. Applied and Environmental Microbiology.
89. Yang B, Wang Y, Qian P-Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135.
90. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, Ludwig W, Glöckner FO, Rosselló-Móra R. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology.
91. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS. 2009. Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. Applied and Environmental Microbiology.
92. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. 2012. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One 7:e33865.

The Accuracy of 16S rRNA Phylogenies

CHAPTER TWO

Ian R. Humphreys

Abstract

The majority of microbiome studies rely on 16S rRNA sequences to assess community diversity and frequently leverage phylogenetic measures to do so. However, there has been relatively little investigation of the accuracy of phylogenies assembled from the rather short and voluminous sequences produced during most microbiome investigations. We developed a statistical simulation framework to quantify the accuracy of such phylogenies and discern the effect of their error. Our software framework subsamples and trims full-length 16S rRNA sequences from 16S rRNA databases to simulate short reads obtained from environmental amplicon sequencing. We then compute similarity and dissimilarity metrics between 16S rRNA phylogenetic trees constructed using short sequences versus full-length sequences to measure the accuracy of phylogenies with differing sequencing parameters. We find that as sequence length increases, the phylogenetic error of truncated sequence phylogenies decreases. We demonstrate that including full-length reference sequences from a database ameliorates this error. Additionally, we find that the phylogenetic diversity within microbial communities and the number of lineages within a phylogeny influences the ability of phylogenies assembled from truncated sequences to accurately recapitulate their full-length counterparts. Furthermore, we present an automated software pipeline that researchers can use to produce phylogenies with optimized accuracy given the results of our simulations. Here, we provide evidence that sequence length and sample diversity drive patterns of similarity between phylogenies constructed using full-length sequences relative to short sequences. Collectively, our findings highlight the importance of experimental design and methodological selection by demonstrating their impact on

phylogenetic tree structure which may ultimately skew the interpretation of phylogeny-derived inferences.

Introduction

The advent of culture-independent genome sequencing has transformed microbiology and created numerous multi-disciplinary fields dedicated to understanding the petabytes of genetic code. In an attempt to interpret these data, microbial sequences are often analyzed on the basis of relative taxonomic abundance and distribution. The 16S rRNA gene encodes the 30S ribosomal protein which constitutes the small-subunit of all bacterial and archaeal ribosomes and provides a universal marker to infer taxa. 16S ribosomal function is conserved thereby constraining the rate and location of tolerable mutations; yet, due to the transcript structure, the 16S gene can be deconstructed into nine hypervariable regions flanked by nine highly conserved regions (Woese et al. 1980 and Yang et al. 2016). As a result, PCR primers can be designed to target highly conserved regions and amplify the intervening hypervariable regions to infer the taxa of bacteria (Van de Peer et al. 1996; Baker et al. 2003).

Since the development of rapid 16S rRNA gene amplicon sequencing (Lane et al. 1985), phylogenetic trees reconstructed using rRNA sequences have been used to describe microbial diversity. In addition to applications in microbial ecology (eg. Lozupone and Knight 2007) and epidemiology (Clarridge 2004), the use of small-subunit RNA sequences is deeply rooted in taxonomy and formed the basis for defining the three domains of life as Archaea, Bacteria, and Eucarya (Woese et al. 1990). While 16S rRNA sequences are used to impute taxonomic classification, these sequences can also be used to construct phylogenies which offer greater insight into microbial function and evolutionary relationships (Lozupone and Knight 2005; Gaulke et al. 2018; Washburne et al. 2018).

Technological limitations and budgetary constraints have resulted in routine use of short sequences obtained from next generation high-throughput sequencing platforms to create phylogenetic trees. Early work demonstrated that phylogenies constructed on the basis of partial 16S rRNA sequences were topologically equivalent to those derived from full-length 16S rRNA sequences (McCarroll et al. 1983), which have been shown to generally correlate with genome phylogenies (Snel et al. 1999). However, more recent studies have demonstrated that the accuracy of phylogenies is affected by the gene(s) used (Rosenberg and Kumar 2001; Case et al. 2006; Wu et al. 2013), the specific region of gene(s) used (Martin et al. 1995; Schloss 2010; Ragan-Kelley et al. 2013; Yang et al. 2016), sequence length (Martin et al. 1995; Graybeal et al. 1998; Rosenburg and Kumar 2001, 2003; Ragan-Kelley et al. 2013), alignment methodology (Lake 1991; Morrison and Ellis 1997; Hall 2005; Ogden and Rosenberg 2006; Schloss 2010), tree reconstruction method (Rosenberg and Kumar 2001; Ogden and Rosenberg 2006), and the genetic diversity of the sequences being compared (reviewed in Nabhan and Sarker 2012). As a result, it is imperative to fully understand how these experimental parameters bias 16S rRNA gene sequence phylogenies that subsequently influence the outcome of phylogenetic-based approaches to study microbial community ecology, such as ClaaTU (Gaulke et al. 2018), UniFrac (Lozupone and Knight 2005; Lozupone et al. 2007), PGLS (Grafen, 1989), Phylofactorization (Washburne et al. 2017), PhILR (Silverman et al. 2017), and PhyloAssigner (Vergin et al. 2013).

To quantitatively investigate how sequencing parameters in environmental 16S studies affect phylogenetic tree structure, and by extension downstream inferences, we established a software framework to simulate short read amplicon sequencing. Using this

framework, we conducted simulations to measure error in phylogenetic trees derived from sequence length, hypervariable region, and sequence diversity. Further, we offer cost-effective suggestions that can improve the ability of short sequences to represent full-length 16S evolutionary history.

Methods

Sequence Selection

We simulated sequencing data that is representative of processed reads generated from 16S rRNA amplicon sequencing studies which are commonly used in the reconstruction of phylogenetic trees using the pre-aligned full-length 16S rRNA database: SILVA 16S LTP version 123 (Yarza et al. 2008). Sequences in the database were filtered based on nucleotide length from primer mapping locations to remove sequences that were too short for our analyses. For all phylogenies, we selected "reference" sequences prior to sampling lineages. Briefly, reference lineages are full-length sequences that provide additional diversity that are used during the reconstruction of phylogenies and are removed prior to analysis. Sequences were selected for references and samples separately according to three sampling paradigms: (1) maximum phylogenetic distance, (2) randomly, and (3) representative taxa based on literature. Maximum phylogenetic distance sampling selects the most distant lineages based off of the full-length all sequence database tree (Wu et al. 2009). Random sampling leverages Perl module, List::Util qw(shuffle) to randomly select sequences from the SILVA 16S database. Representative taxa sampling applies filtering criterion on random sampling to exclude taxa that are not present within a desired taxonomic level.

Simulation Methods

We investigated how sequence length, position within the 16S rRNA gene, taxonomic diversity, sample size, and number of full-length reference lineages effect measures of phylogenetic accuracy. From the selected aligned full-length 16S rRNA

sequences, we simulated reads obtained from high-throughput DNA sequencing platforms by cutting the length of sequences based on the alignment positions that correspond to PCR primers from literature in the 5' to 3' direction (Table 1). Sample diversity was integrated into the analyses based on the sampling paradigms listed above. The maximum number of reference sequences were determined prior to each experiment and reserved from the potential sampling pool. References were removed after reconstructing complete phylogenies using both simulated short read sequences or full-length sequences and full-length references with R package ape::drop.tip to produce reference-guided phylogenies.

Construction of Phylogenetic Trees

For each test condition, we constructed four phylogenetic trees: full-length (FL) and short read (SR) phylogenies that are comprised of only sampled sequences and reference-guided full-length (RGFL) and reference-guided short read (RGSR) phylogenies that include full-length reference sequences that were pruned out prior to analysis. All phylogenetic trees were reconstructed using FastTree-2.1.10 -nt -gtr to specify generalized time-reversible model of nucleotide substitution (Price et al. 2010).

Phylogenetic Statistical Analyses

All analyses were conducted using R (R Core Team 2018). To quantify the differences in phylogenetic diversity between phylogenetic trees reconstructed under different conditions, we computed tip-to-tip distances using the base R function, cophenetic on each newick format phylogenetic tree. This produces a distance matrix

between each tip within a tree by summing the total branch lengths that segregate each pair of lineages. To compare trees (FL, SR, RGFL, and RGSR), we correlate these distance matrices with a mantel correlation using vegan::mantel, which is interpreted such that higher correlation coefficients correspond to more similar phylogenies. In the case of comparisons with FL trees, these correlation coefficients serve as a measure of accuracy. Tree topology was quantified using a normalized Robinson-Foulds distance metric using phangorn::RF.dist in a pairwise manner between phylogenetic trees with the same sampled sequences (FL, SR, RGFL, and RGSR). Robinson-Foulds distances assess the structural differences between two trees by computing the number of partitions that differ between trees (Robinson and Foulds 1981). When normalized Robinson-Foulds distances are calculated between FL trees and alternative phylogenies, a lower value corresponds to higher accuracy.

Results

Increased sequence length improves phylogenetic accuracy

We developed and applied a software framework which subsamples and simulates short read 16S rRNA sequences from pre-aligned databases to assess the accuracy of phylogenetic trees reconstructed from reads typically obtained during environmental amplicon sequencing experiments. Briefly, the starting locations of simulated short reads were identified by the location of universal primers in an aligned *E. coli* sequence. Sequences were then trimmed to a selected number of nucleotides. We randomly sampled 1000 sequences from the SILVA 16S LTP database (Yarza et al. 2008) and trimmed sequences to 100, 200, 300, 400, and 500 nucleotides from the beginning of the V2, V3, V4, V5, and V6 hyper-variable regions by mapping the nucleotide alignment position of PCR primers in the 5' to 3' direction (Table 1). We reconstructed phylogenetic trees from these simulated short reads (SR) and compared their phylogenetic diversity and topology to full-length (FL) phylogenies of the same sequences to assess overall phylogenetic accuracy.

We found that as sequence length increases, irrespective to hyper-variable location within the 16S gene, measures of both phylogenetic diversity and topology become more similar to the FL phylogeny (Figure 2A,C). The greatest improvement to the accuracy of SR phylogenetic diversity compared to FL phylogenetic diversity occurred between 100 and 200 nucleotides with an average increase to the mean correlation across all hypervariable regions of 6.62×10^{-4} per additional nucleotide and secondarily between 200 and 300 nucleotide long sequences with an average increase of 5.02×10^{-4} per additional nucleotide. The accuracy of tree topology improved at a

steadier rate across sequence lengths. These results corroborate prior work that demonstrates increased sequence length yields more accurate phylogenies (Graybeal 1998; Rosenberg and Kumar 2003; Schloss 2010; Ragan-Kelley et al. 2013) through both phylogenetic diversity and topology while contextualizing this finding to 16S gene sequence phylogenies.

V4-hypervariable region most closely reconstructs full-length 16S phylogeny

We found that given highly diverse phylogenies reconstructed with between 100 to 400 nucleotide SR sequences, the V4 hypervariable region generally best reconstructs FL phylogenetic diversity and topology followed by V2, V3, V5, and V6 (as in Table 1) (Figure 2A,C). At 100 nucleotide SR phylogenies, the V3 and V4 hypervariable regions did not display significantly different phylogenetic diversities compared to FL phylogenies (Tukey HSDT on one-way ANOVA; $F_{4,495} = 678$, $P > 0.05$). However, the mean V4 topological accuracy was significantly greater than V3 (Tukey HSDT on one-way ANOVA; $F_{4,495} = 1023$, $P < 0.001$). The effect of a hypervariable region was minimal on phylogenetic diversity at 500 nucleotide long SR phylogenies. Only four of the ten pairwise comparisons between hypervariable regions (V1-V5, V3-V5, V4-V5, and V5-V6) had statistically significant differences in the mean phylogenetic diversity correlation of 500 nucleotide SR to FL phylogenies (Tukey HSDT on one-way ANOVA; $F_{4,495} = 19.62$, $P < 0.001$). These observations support prior assertions that for short reads, the V4 hypervariable region describes the greatest discriminatory power (Schloss 2010; Ragan-Kelley et al. 2013; Yang et al. 2016) and that gene-regions have a greater effect on phylogenetic inferences than sequence length (Martin et al. 1995). However, by

conducting statistical comparisons with both phylogenetic distance and topology metrics, we demonstrate that SR phylogenies reconstructed from a hypervariable region may inadequately recapitulate the phylogenetic diversity of FL sequences yet be highly correlated with FL topology (eg. V2).

Reference sequences improve phylogenetic accuracy

Prior work has demonstrated that the phylogenetic diversity within samples and total sample size influence phylogenetic accuracy (reviewed in Nabhan and Sarker 2012). Incorporating additional sequences into the reconstruction of phylogenetic trees and then removing the sequences has been shown to reduce phylogenetic error (Rosenberg and Kumar 2001; Pollock et al. 2002) and has been used when answering biological questions about microbial community diversity (Sharpton et al. 2011; Riesenfeld and Pollard 2013; O'Dwyer et al. 2015; Gaulke et al. 2018). We extended this concept of adding sequences from a full-length 16S database during phylogenetic tree reconstruction into our analyses to determine how the addition of full-length sequences and their subsequent removal prior to analysis influences phylogenetic accuracy. We refer to these full-length sequences as *reference sequences*, and phylogenetic trees that incorporate reference sequences as reference-guided full-length (RGFL) and reference-guided short read (RGSR) phylogenies.

We selected 1000 full-length 16S reference sequences from the SILVA LTP database that represent taxa with maximum phylogenetic distance between them. These reference sequences were added to the 1000 randomly selected sample 16S gene sequences and used to reconstruct phylogenies. We found statistically significant

improvement in the accuracy of phylogenetic diversity when 1000 reference sequences are included during phylogenetic tree reconstruction (mean phylogenetic diversity correlation to FL phylogenies is 0.054 higher in RGSR than SR phylogenies; paired t-test $p < 0.001$). Less pronounced but still significant improvement in topology accuracy for RGSR phylogenies (mean Robinson-Foulds distance to FL phylogenies is 0.016 lower in RGSR than SR phylogenies; paired t-test $p < 0.001$) was also observed (Figure 2B,D). Additionally, including reference sequences greatly reduced the effect of sequence length and hypervariable region on phylogenetic diversity in comparisons between RGSR and FL phylogenies.

Next, we investigated how the quantity of full-length reference sequences affects phylogenetic diversity and topology. We reconstructed phylogenies with 1000 sample sequences and compared 100-nucleotide long RGSR phylogenies that contain anywhere from 1 to 5000 reference sequences. The resulting 5000 different phylogenies were each compared to FL phylogenies of the 1000 sampled sequences. As before, all sampled sequences and references sequences were selected based on maximum phylogenetic distance. We found that there was a sharp decline in the rate of change in the similarity of phylogenetic diversity between RGSR and FL phylogenies at approximately 750 reference sequences. Adding additional reference sequences beyond 1000 to the phylogeny yielded little additional effect on diversity as compared to phylogenies containing only 1000 reference sequences (Supplemental Figure 1).

Variation along the 16S rRNA gene defines regions that yield accurate phylogenies and regions that reduce accuracy

We quantified the accuracy of phylogenies reconstructed from short reads that represent every continuous 100, 300, and 500 nucleotide stretch within the 16S gene to determine the effect of the per-nucleotide variation in highly conserved and hyper variable regions within the 16S rRNA gene. To conduct this analysis, we reconstructed phylogenies with 100, 300, and 500 nucleotide long SRs that began at each nucleotide position in the 6888 position SILVA LTP alignment in the 5' to 3' direction and compared each SR phylogeny to the corresponding FL phylogeny. We simulated 1000 short reads and 1000 reference sequences which were selected to represent taxa that maximize the phylogenetic distance between each pair of sequences. Our results showed that phylogenies reconstructed with SRs that include regions of the 16S gene that are adjacent to the alignment position where the final nucleotide of PCR primers map to (eg. 515F and 784F alignment positions 3281 and 3800 respectively) result in decreased phylogenetic accuracy than SRs that include portions of highly conserved regions compared to FL phylogenies (Figure 3). These results reinforced our prior findings that the inclusion of reference sequences alongside SR improves the accuracy of phylogenetic diversity compared to FL phylogenies (Figure 3B). Additionally, we demonstrated that increased sequence length compensates for regions of poor phylogenic signal (Figure 3).

Microbial taxonomic diversity contributes to phylogenetic accuracy

To assess how community diversity affects phylogenetic accuracy, we sampled sequences using taxonomically biased and unbiased approaches and measured the effect of this bias on phylogenetic accuracy. Specifically, we sample sequences that represent taxa with maximum phylogenetic distance between each sequence, were randomly

distributed across the LTP phylogeny, or restricted to taxa that are all annotated in a single phylum or order. Simulated short read sequences were trimmed to 100-600 nucleotides from the beginning of the V4 hypervariable region in the 5' to 3' direction (Table 1). 1000 reference sequences were selected based on maximizing the phylogenetic distance between each sequence and combined with the simulated short read sequences to reconstruct 500 tip phylogenies. The four single phyla-level phylogenies contained only taxa within the Proteobacteria, Bacteroidetes, Actinobacteria, and Firmicutes phyla and the three order-level phylogenies contained only taxa within the Bacillales, Flavobacteriales, and Actinobacteridae orders. Randomly sampled sequences may have included up to 35 different phyla and maximum diversity samples were always comprised of the same 20 phyla. We found that phylogenetic trees reconstructed using taxonomically constrained sampling resulted in decreased accuracy as measured by the phylogenetic diversity and topology of SR phylogenies compared to FL phylogenies (Figure 4). Additionally, this effect was magnified at shorter read lengths. For example, the mean phylogenetic diversity correlation coefficient between 100 nucleotide SR and FL phylogenies reconstructed with random sequences compared to single order phylogenies decreased by 0.243 compared to a 0.169 decrease at 600 nucleotides (Tukey HSDT on one-way ANOVA; 100nucletoide $F_{3,70} = 85.55$, $P < 0.001$; 600 nucleotide $F_{3,70} = 29.75$, $P < 0.001$). When including 1000 reference sequences selected to represent taxa with maximum phylogenetic distance between each sequence in tree reconstruction, we saw that the phylogenetic diversity of RGSR phylogenies improved compared to FL sequences, however, the effect of reference sequences was modest on topological accuracy. We hypothesize that our observation of phylogenies with lower taxonomic

diversity that result in less accurate phylogenies compared to full-length sequences will extend to phylogenies reconstructed from microbial communities with low phylogenetic diversity.

Large phylogenies are more sensitive to topological errors

We used 100, 500, 1000, and 7500 sequences to reconstruct phylogenies and test the effect of number of taxa on the accuracy of phylogenies. We saw that as the number of sequences that comprise a phylogeny increases, the number of sequences does not have an ordered effect on the phylogenetic diversity of SR compared to FL phylogenies (Figure 5A). However the addition of 1000 reference sequences generally introduced structure to the effect of phylogeny size in the calculation of phylogenies diversity for RGSR phylogenies compared to FL phylogenies (Figure 5B). We did find that phylogenies with fewer number of sequences resulted in improved topology for SR and RGSR compared to FL phylogenies (Figure 5C,D). It is difficult to disentangle the effect of the number of sequences from phylogenetic distance between taxa on the accuracy of phylogenies because we applied our maximum phylogenetic distance sequence selection methodology which selects sequences based on maximum phylogenetic distance from each other. Consequently, we show that increasing the number of sequences in a phylogeny, and thus decreasing the phylogenetic distance between neighboring tips in the tree, increased topological and phylogenetic diversity error in SR phylogenies compared to FL phylogenies.

Discussion

Reconstructing accurate phylogenies is an exciting prospect for strict systematists and general microbiologists alike because they enable researchers to glean insight into the evolutionary history of organisms and speculate about their functional potential. We offer new insight into methods for improving phylogenetic accuracy within the context of microbiome studies. Prior studies rooted in theory and simulation that attempted to quantify phylogenetic accuracy were conducted in the 1990's and early 2000's but lacked validation on 16S rRNA data. Many of the parameters that affect phylogenies originally outlined in these studies have remained under appreciated in microbiome sciences. Our work exemplifies some of these difficulties in phylogenetic reconstruction and subsequently phylogeny-informed analyses in part due to the high per-base genetic variation within 16S rRNA genes that make it a strong phylogenetic marker (Woese et al. 1980; Noller and Woese 1981; Ashelford et al. 2005).

The simulation-based experimental design we employ to assess the ability of short read 16S gene sequences to reconstruct the full-length 16S gene phylogeny mitigates the effects of poor sequence quality and alignment by leveraging SILVA's hand-curated 16S sequence database. That said, while these sequences are of high quality (Schloss 2009; Schloss 2010), they may still contain some errors despite database curation and thereby not necessarily reflect the evolutionary relationships obtained through phylogenomic analysis. We briefly considered the effect of alignment on phylogenetic accuracy however prior work already demonstrates the effects of lane masking (Lane 1991; Schloss 2010) and alignment quality of databases on phylogenies (Schloss 2010). As a result, we decided to limit our analyses to SILVA LTP 16S rRNA database sequences

due to their superior sequence quality and alignment that better estimate measures of phylogenetic diversity (Schloss 2009; 2010). By using a single database, we ensure that additional variation to sequences that are introduced during curation is negated despite reducing the maximum size of phylogenies and sequence diversity. We understand that the present study does not exhaustively address all sources of methodological errors that may influence phylogenetic analyses and does contain limitations as a result of simulations. For example, our simulation study design restricts our ability to extrapolate which hypervariable region primers should be used during PCR to best represent global bacterial phylogenies. We do however show that the V4 hypervariable region, when detectable, produces the most similar phylogeny compared to FL sequences within the SILVA 16S database. Additionally, our application of constrained taxonomic sampling still allowed us to ascertain effect of sample diversity on the accuracy of SR phylogenetic diversity without biological samples.

Using simulated collections of short reads, we demonstrated that phylogenetic trees comprised of sequences that represent taxa with maximum phylogenetic distances between each pair of tips produce more accurate phylogenies than lower phylogenetic distance phylogenies compared to the full-length sequences. We hypothesize that phylogenies with low phylogenetic distance between taxa are less likely to accurately infer correct branch length and positional accuracy than phylogenies reconstructed with highly diverse taxa. We had initially hypothesized that the addition of full-length reference sequences would reduce the effects of sequence similarity on the ability to infer accurate phylogenies by artificially inflating diversity during tree reconstruction; however the composition of reference sequences may play a pivotal role in the strength of

this effect. Prior work which showed that introducing additional sequences that bisect long branches improves the accuracy of branch lengths (Hillis 1998) and consequently phylogenetic diversity supports this hypothesis. Similarly, as the number of reference sequences is increased, the rate of improvement in similarity between RGSR and FL phylogenies decreases. This indicates that there may be scenarios in which the proportion of additional phylogenetic diversity that is achieved by adding a new reference sequence yields negligible improvements in phylogenetic accuracy. These findings corroborate prior work which finds that the benefit of RGFL phylogenies are constrained by the phylogenetic diversity captured within the references (Sharpton et al. 2011).

As we move towards improved sequencing technologies that enable rapid and accurate sequencing nearing full-length 16S gene segments we expect the incorporation of high-quality reference sequences will continue to improve phylogenetic accuracy. Regardless of sequencing quality and length, the added phylogenetic diversity that provides additional contextualization for lineages may improve the resolution of phylogenetic trees. Our results comparing FL to RGFL (data not shown) indicated that there are still differences between both phylogenetic diversity and tree topology. We hypothesize that these differences between FL and RGFL phylogenies provide a further gain in accuracy due to the additional information in the form of added phylogenetic diversity with reference sequences. Further, we extrapolate that multi-gene, protein, or genome phylogenies would benefit from pertinent reference sequences to guide the placement of the samples.

In summary, we have provided evidence that sequence length, sample size, and sample diversity drive patterns of accuracy between phylogenies reconstructed using full-

length and short sequences. We demonstrated that reference sequences improve the precision of short sequences to reconstruct 16S rRNA relationships between taxa. As a result, we recommend that researchers integrate full-length reference sequences into their phylogenetic reconstruction methods to improve the accuracy of their phylogenies. We found that generally, the greatest increase in phylogenetic accuracy occurs between 100 and 200 nucleotides, and therefore suggest that 16S rRNA based studies should strive to be conducted with reads of at least 200 nucleotides in length. However, when conducting analyses that apply clade-based phylogenetic tools, we recommend longer read lengths due to the continuous improvement in topological accuracy compared to full-length phylogenies. Furthermore, when reconstructing phylogenies with closely related organisms, we recommend longer sequence lengths to minimize phylogenetic error. Together, our findings have highlighted the importance of experimental design and methodological selection by demonstrating their impact on phylogenetic tree structure which may ultimately skew the interpretation of phylogeny-driven inferences.

References

1. Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L. 2008. Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. PLoS ONE.
2. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71:7724–7736.
3. Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555.
4. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. PNAS.
5. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. 2007. Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. Applied and Environmental Microbiology.
6. Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of Microbiological Methods.
7. Clarridge JE. 2004. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews.
8. Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. 2018. Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota. mBio.
9. Grafen A. 1989. The Phylogenetic Regression. Philosophical Transactions of the Royal Society B: Biological Sciences.
10. Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47:9–17.
11. Hall BG. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol 22:792–802.
12. Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 47:3–8.
13. Lake JA. 1991. The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol 8:378–385.
14. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences.
15. Lane DJ. 1991. 16S/23S rRNA Sequencing. Nucleic Acid Techniques in bacterial Systematics. New York: Wiley. 115-175.
16. Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology.
17. Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences.

18. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology.
19. Martin MJ, González-Candelas F, Sobrino F, Dopazo J. 1995. A method for determining the position and size of optimal sequence regions for phylogenetic analysis. J Mol Evol 41:1128–1138.
20. McCarroll R, Olsen GJ, Stahl YD, Woese CR, Sogin ML. 1983. Nucleotide sequence of the Dictyostelium discoideum small-subunit ribosomal ribonucleic acid inferred from the gene sequence: evolutionary implications. Biochemistry.
21. Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol 14:428–441.
22. Nabhan AR, Sarkar IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics.
23. Noller HF, Woese CR. 1981. Secondary structure of 16S ribosomal RNA. Science 212:403–411.
24. O'Dwyer JP, Kembel SW, Sharpton TJ. 2015. Backbones of evolutionary history test biodiversity theory for microbes. Proceedings of the National Academy of Sciences.
25. Ogden TH, Heath Ogden T, Rosenberg MS. 2006. Multiple Sequence Alignment Accuracy and Phylogenetic Inference. Systematic Biology.
26. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. Systematic Biology.
27. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE.
28. R Core Team. 2018. R: A language and environment for statistical computing. R Foundation 505 for Statistical Computing, Vienna, Austria.
29. Ragan-Kelley B, Walters WA, McDonald D, Riley J, Granger BE, Gonzalez A, Knight R, Perez F, Gregory Caporaso J. 2013. Collaborative cloud-enabled tools allow rapid, reproducible biological insights. The ISME Journal.
30. Riesenfeld SJ, Pollard KS. 2013. Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis. BMC Genomics.
31. Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences.
32. Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proceedings of the National Academy of Sciences.
33. Rosenberg MS, Kumar S. 2003. Taxon Sampling, Bioinformatics, and Phylogenomics. Systematic Biology.
34. Schloss PD. 2009. A high-throughput DNA sequence aligner for microbial ecology studies. PLoS One 4:e8230.
35. Schloss PD. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6:e1000844.

36. Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA, Pollard KS. 2011. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data. PLoS Computational Biology.
37. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. eLife.
38. Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nature Genetics.
39. Turner S, Pryer KM, Miao VPW, Palmer JD. 1999. Investigating Deep Phylogenetic Relationships among Cyanobacteria and Plastids by Small Subunit rRNA Sequence Analysis. The Journal of Eukaryotic Microbiology.
40. Van de Peer Y, Chapelle S, De Wachter R. 1996. A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Research.
41. Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, Kilpert F, Worden AZ, Giovannoni SJ. 2013. High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. The ISME Journal.
42. Wang Y, Qian P-Y. 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One 4:e7401.
43. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ.
44. Washburne AD, Morton JT, Sanders J, McDonald D, Zhu Q, Oliverio AM, Knight R. 2018. Methods for phylogenetic analysis of microbiome data. Nat Microbiol 3:652–661.
45. Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius R, Gutell R, Hogan JJ, Noller HF. 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Research.
46. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences.
47. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature.
48. Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. PLoS ONE.
49. Yang B, Wang Y, Qian P-Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135.

50. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, Ludwig W, Glöckner FO, Rosselló-Móra R. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology.

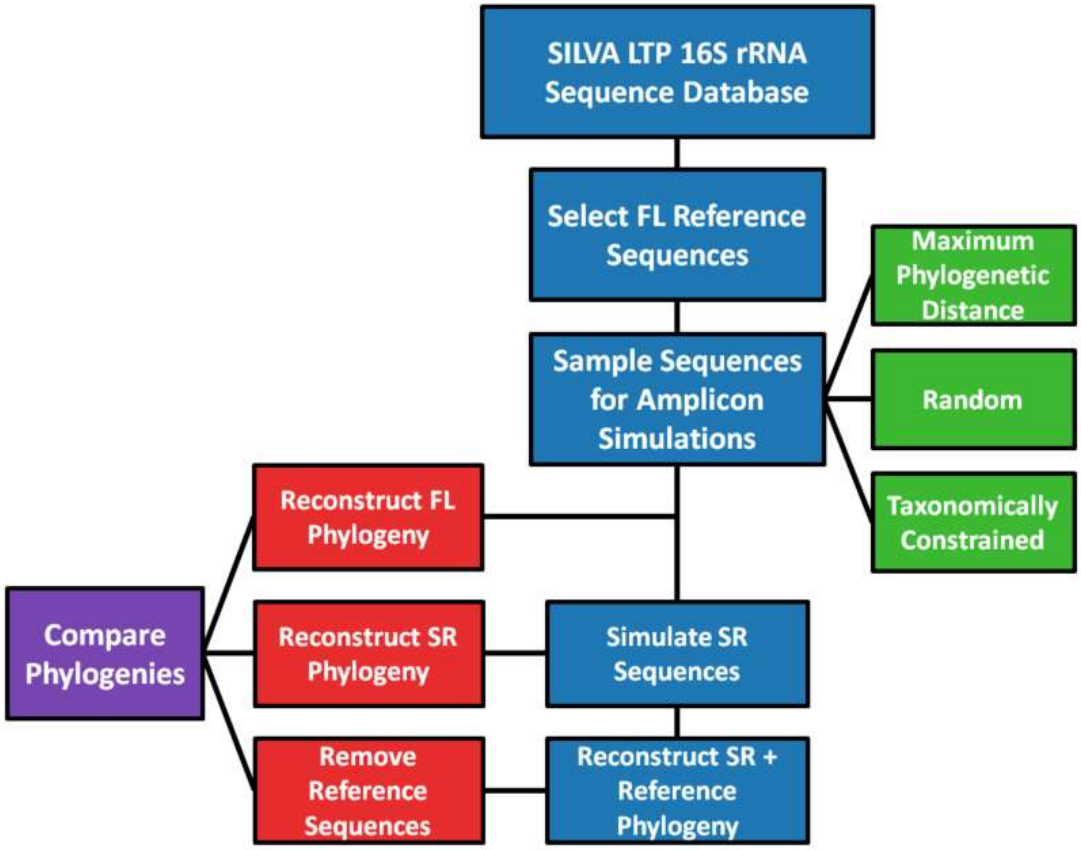

Figure 2.1 | **Simulation Framework Overview**

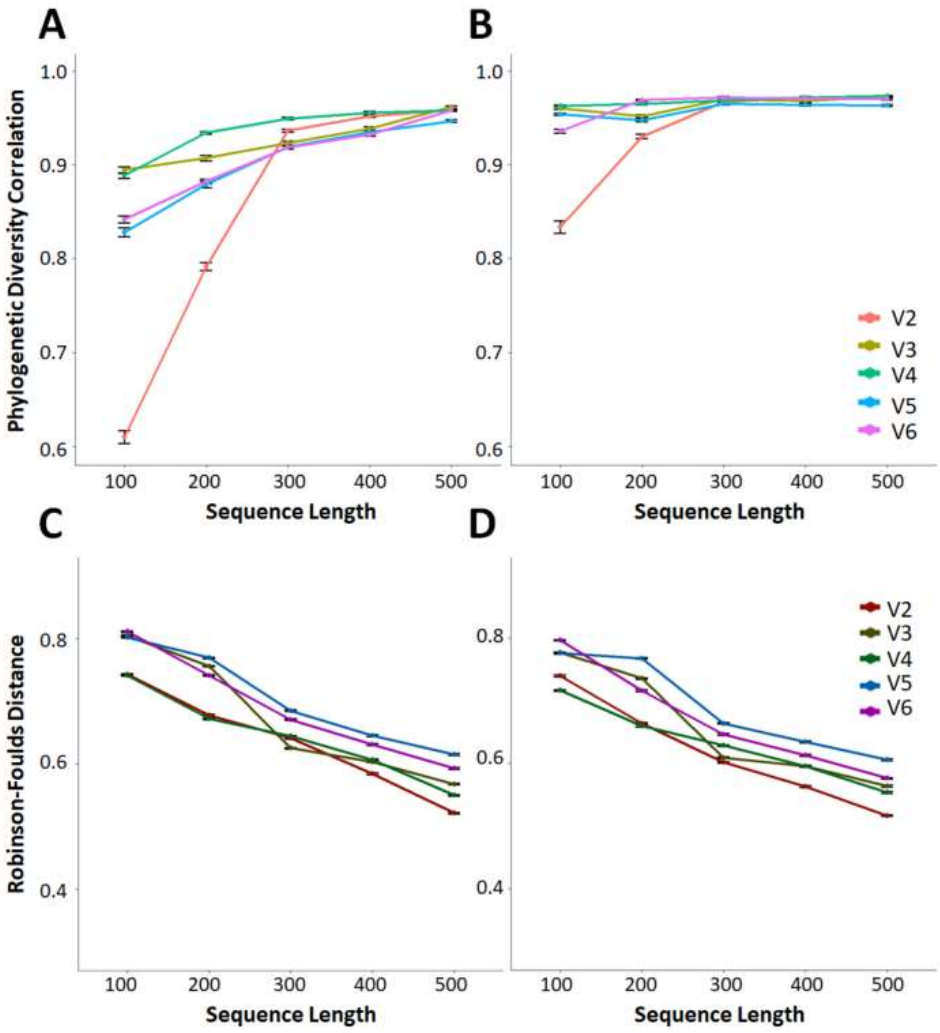

Figure 2.2 | **The accuracy of phylogenetic diversity and tree topology across read length and hypervariable regions.** Each point represents the mean value from 100 comparisons between two phylogenetic trees, one reconstructed with SR or RGSR and the other with FL sequences. Phylogenies were reconstructed using 1000 randomly selected sequences from SILVA 16S LTP database after filtering out sequences shorter than 550 nucleotides from the beginning of the V6 primer. SR and RGSR sequences were trimmed to 100, 200, 300, 400, and 500 nucleotides from the beginning of the primer location within the SINA aligned sequences for hypervariable regions V2-V6 in the 5' to 3' direction. (A) Phylogenetic diversity of SR phylogenies compared to FL phylogenies. (B) Phylogenetic diversity of RGSR phylogenies compared to FL phylogenies. (A,B) The mean mantel correlation between pairwise tip-to-tip distance matrices are reported with standard error bars. (C) Topological dissimilarity of SR phylogenies compared to FL phylogenies. (D) Topological dissimilarity of RGSR phylogenies compared to FL phylogenies. (C,D) The mean value of normalized Robinson-Foulds distance metric of tree topology between phylogenies is reported with standard error bars.

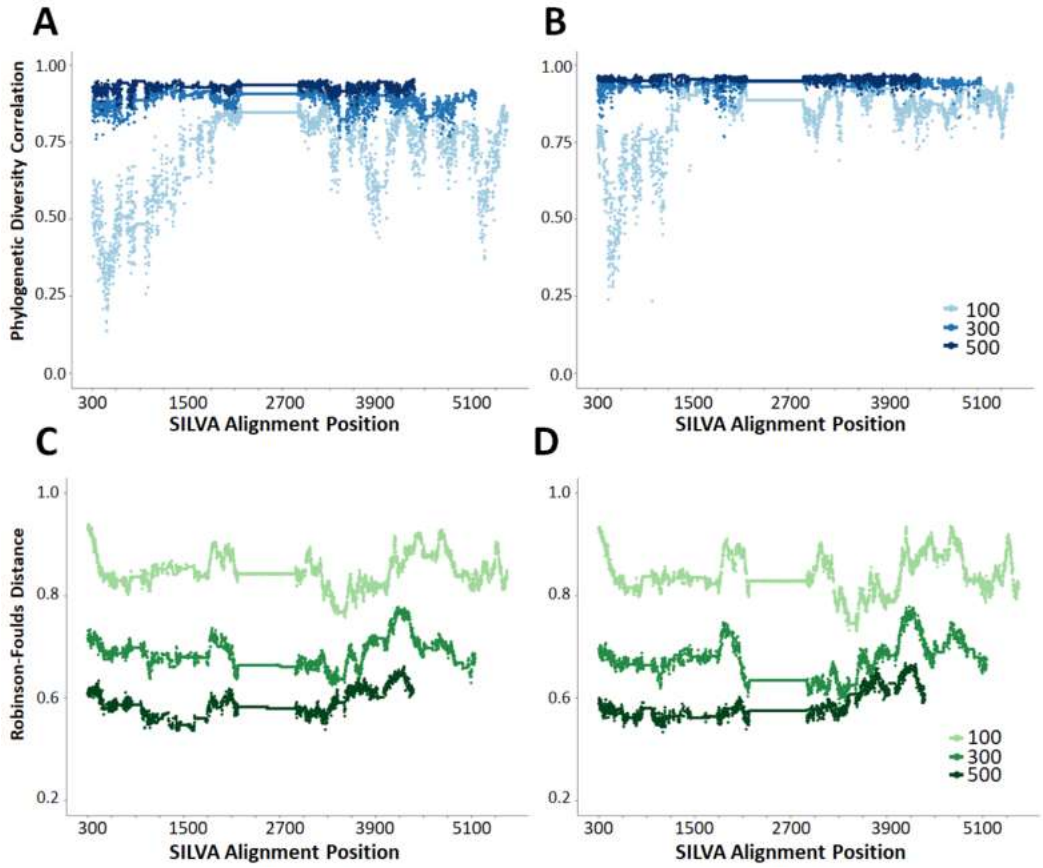

Figure 2.3 | **A per-base sweep of phylogenetic accuracy across SILVA 16S rRNA alignment.** Each point is a comparison between two phylogenetic trees, one reconstructed with SR or RGSR and the other with FL sequences. Phylogenies were reconstructed using 1000 sequences from SILVA 16S LTP database after filtering out sequences shorter than 550 nucleotides from the beginning of the V6 primer location. The sampled sequences were selected to maximize the phylogenetic distance between each two tips. At each position in the SILVA 6888 column alignment in the 5' to 3' direction, SR and RGSR were trimmed to 100, 300, and 600 nucleotides. A) Phylogenetic diversity of SR phylogenies compared to FL phylogenies. (B) Phylogenetic diversity of RGSR phylogenies compared to FL phylogenies. (A,B) Mantel correlation between pairwise tip-to-tip distance matrices. (C) Topological dissimilarity of SR phylogenies compared to FL phylogenies. (D) Topological dissimilarity of RGSR phylogenies compared to FL phylogenies. (C,D) Normalized Robinson-Foulds distance metric of tree topology between phylogenies.

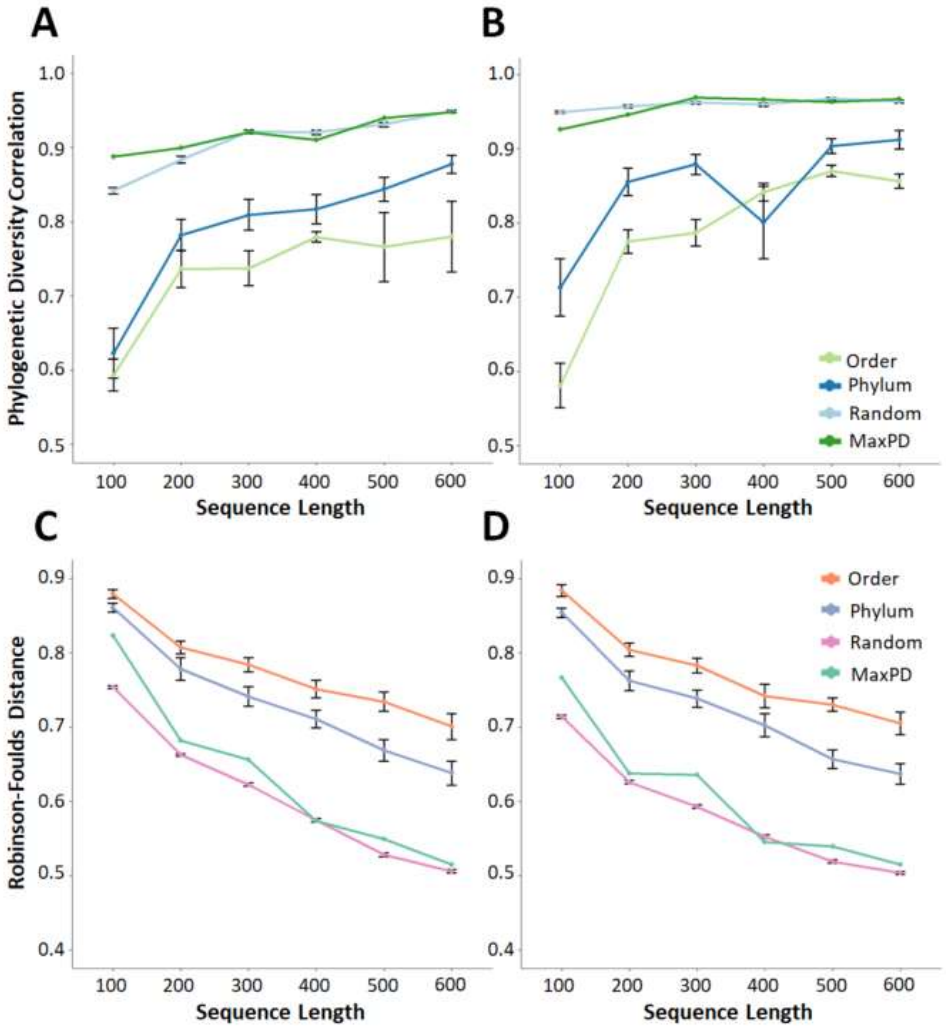

Figure 2.4 | **Phylogenetic diversity and topological accuracy across sampling spaces.** Three types of sampling were employed to select sequences from SILVA 16S LTP: constrained sampling to sequences annotated to 4 selected phyla and 3 orders, random sampling conducted with 100 replicates, and selecting to maximize the phylogenetic distance between each two tips. Phylogenies were reconstructed using 1000 sequences from SILVA 16S LTP database after filtering out sequences shorter than 750 nucleotides from the beginning of the V4 primer location. SR and RGSR sequences were trimmed to 100, 200, 300, 400, 500, and 600 nucleotides from the beginning of the V4 hypervariable region in the 5' to 3' direction. A) Phylogenetic diversity of SR phylogenies compared to FL phylogenies. (B) Phylogenetic diversity of RGSR phylogenies compared to FL phylogenies. (A,B) The mean mantel correlation between pairwise tip-to-tip distance matrices are reported with standard error bars. (C) Topological dissimilarity of SR phylogenies compared to FL phylogenies. (D) Topological dissimilarity of RGSR phylogenies compared to FL phylogenies. (C,D) The mean value of normalized Robinson-Foulds distance metric of tree topology between phylogenies is reported with standard error bars.

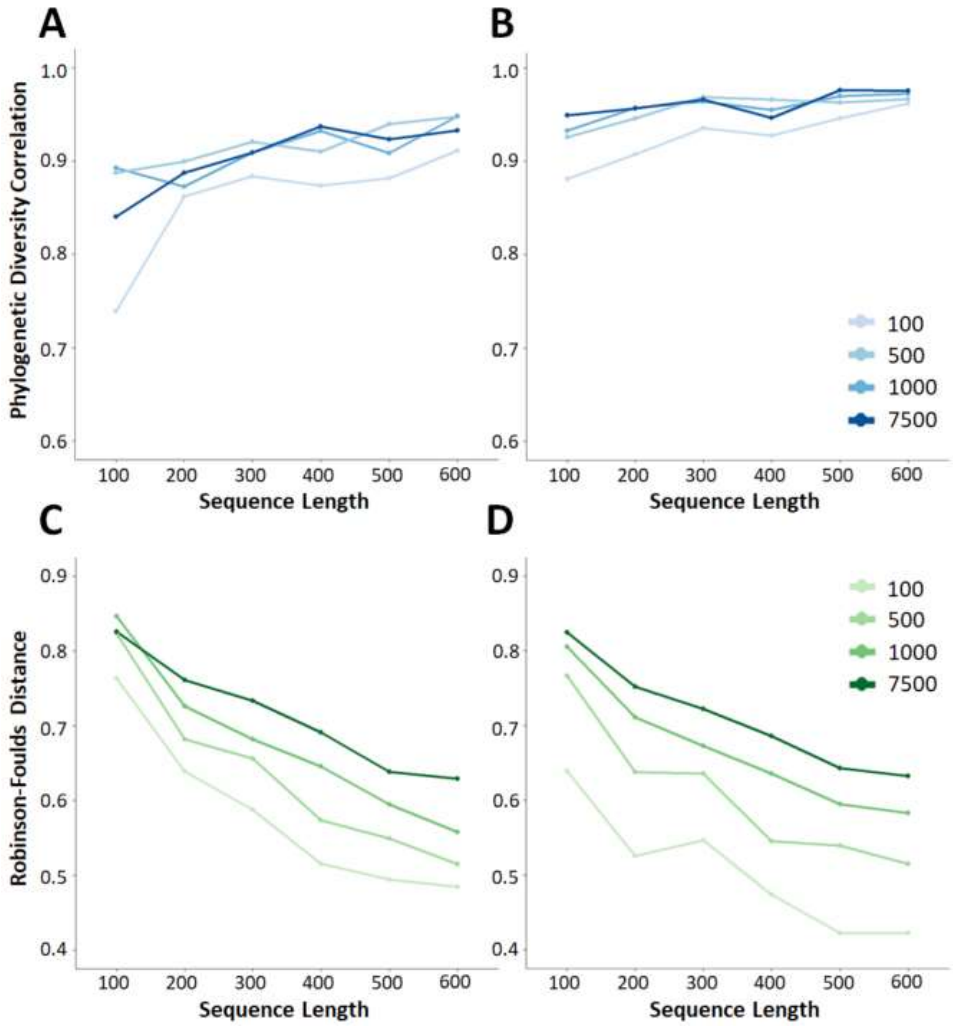

Figure 2.5 | **The effect of sample size on phylogenetic accuracy.** Each point is a comparison between two phylogenetic trees, one reconstructed with SR or RGSR and the other with FL sequences. The 100, 500, 1000, or 5000 sequences in each phylogeny were selected to maximize the phylogenetic distance between each two tips after filtering out sequences shorter than 750 nucleotides from the beginning of the V4 primer. SR and RGSR sequences were trimmed to 100. 200, 300, 400, 500, and 600 nucleotides from the beginning of the V4 hypervariable in the 5' to 3' direction. A) Phylogenetic diversity of SR phylogenies compared to FL phylogenies. (B) Phylogenetic diversity of RGSR phylogenies compared to FL phylogenies. (A,B) Mantel correlation between pairwise tip-to-tip distance matrices are reported for each sequence length and number of lineages. (C) Topological dissimilarity of SR phylogenies compared to FL phylogenies. (D) Topological dissimilarity of RGSR phylogenies compared to FL phylogenies. (C,D) Normalized Robinson-Foulds distance metric of tree topology between phylogenies are reported for each sequence length and number of lineages

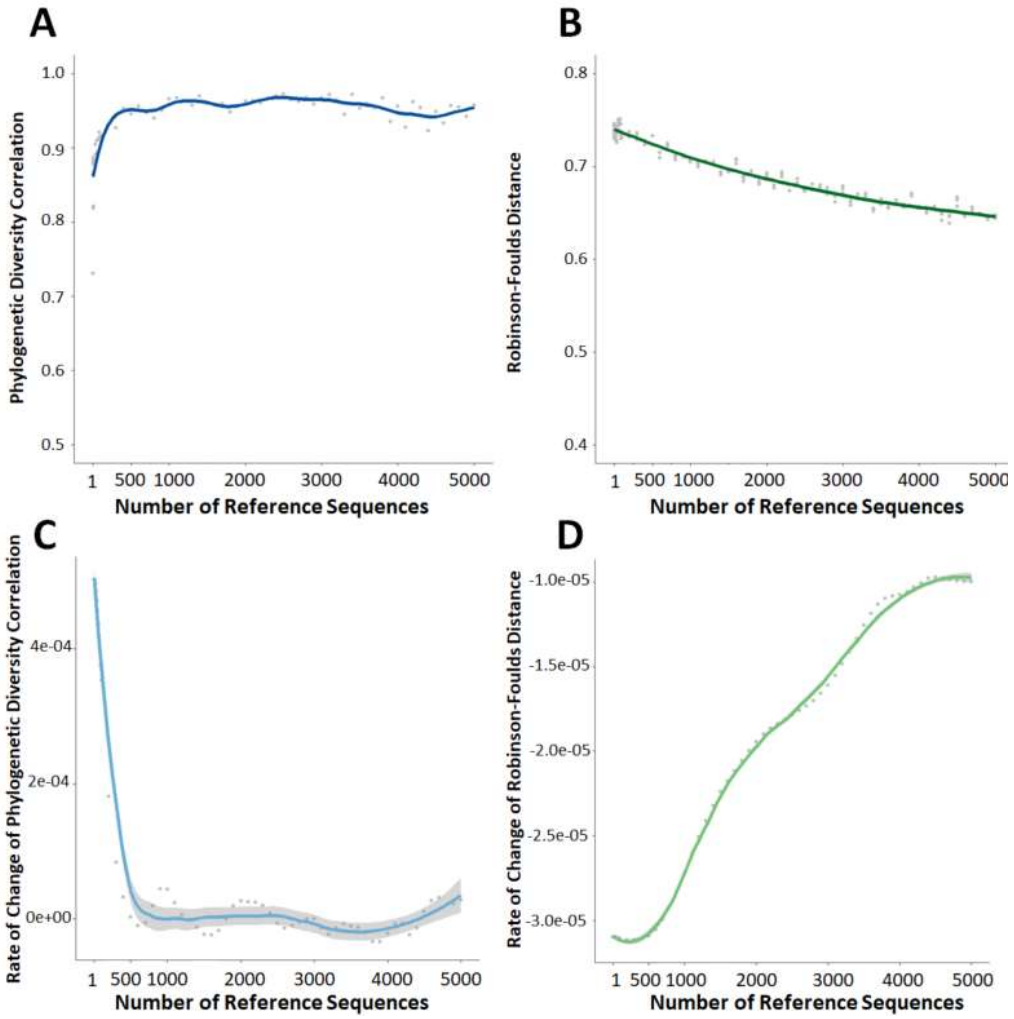

Figure Supplemental 2.1 | **Rate of accuracy improvement with increased reference sequences.** Each point is a comparison between two phylogenetic trees one reconstructed with RGSR and the other with FL sequences. The points are connected with a smooth spline smoothing function for readability. We reserved 5000 sequences that maximize the phylogenetic distance between each two tips from the SILVA LTP 16S sequence database for use as reference sequence's and constructed phylogenetic trees with the next 1000 sequences that maximized the phylogenetic distance. We constructed phylogenies with 1-5000 full-length reference sequences in triplicate and compare the phylogenies reconstructed using RGSR that are 100 nucleotides long from the beginning of the V4 hypervariable region in the 5' to 3' direction. (A) Mantel correlation of tip to tip distances between RGSR and FL. (B) Robinson-Foulds distance metric on RGSR and FL. (C) Rate of change of tip to tip distance correlation. (D) Rate of change of Robinson-Foulds metric. We observed a plateau in the rate of change for both phylogenetic diversity and topology around approximately 750 reference sequences (C,D).

Table 2.1 | **Position of primers in SILVA 16S LTP 6888 Column Alignment** (5’ to 3’)

Region	Position	Primer	Aligned SILVA (E. Coli)	Reference
V2	104F	GGCGVACGGGTGAGTAA	978	Wang and Qian 2009
V3	357F	CTCCTACGGGAGGCAGCAG	2156	Turner et al. 1999
V4	515F	GTGCCAGCMGCCGCGGTAA	3281	Caporaso et al. 2011
V5	784F	AGGATTAGATACCCTGGTA	3800	Andersson et al. 2008
V6	986F	TCGATGCAACGCGAAGAA	4377	Chakravorty et al. 2007

GENERAL CONCLUSIONS

CHAPTER THREE

Ian R. Humphreys

Over the centuries, phylogenetic trees have been reconstructed to compare organisms to one another based on observational phenotypes and ultimately gain insight into evolutionary relationships. However, the advent of genomic sequencing has facilitated rapid generation of nucleic acid based phylogenies which have been extensively applied to microbial systems. The explosive growth of culture-independent based genomic studies in recent decades have resulted in the exponential expansion of known microbial diversity which can be better understood through phylogenetic inferences that contextualize newly discovered microorganisms (Hugenholtz et al. 1998; Hug et al. 2016). Further technological advances have enabled researchers to leverage vast amounts of data obtained through shotgun metagenomics and more recently single-cell sequencing which provide insight into functional potential and fine-scale variation between organisms. Yet 16S rRNA gene based phylogenetic analyses continue to provide an inexpensive method for obtaining valuable information about the composition of microbial communities (Thompson et al. 2017).

Phylogenetic analyses are conducted to integrate information about the evolutionary relationships between organisms in a phylogenetic tree, and because these trees are used to inform a variety of analyses, their accuracy is important to quantify. For example, phylogenies are used to infer community diversity (Lozupone and Knight 2005), reveal patterns of trait selection (Gaulke et al. 2018; Washburne et al. 2018), identify epidemiological trends (Clarridge 2004), and inform taxonomic classification (Yilmaz et al. 2014). Moreover, within group diversity (richness) can be computed using phylogenetic distances between lineages or clades within a tree through summing the total branch lengths, branch lengths between lineages, or number of nodes. When paired

with sample information, a phylogenetic tree that incorporates two groups can be used to assess measures of between group diversity (β-diversity) (eg. UniFrac distance (Lozupone and Knight 2005; Lozupone et al. 2007)). Phylogenies can also be used to map the evolution of trait selection by identifying phylogenetic and co-phylogenetic signal (Gaulke et al. 2018). Additionally, phylogeny can inform taxonomic classification. In 1990, Carl Woese used phylogenetic inferences based on 16S and 18S rRNA sequences to propose three domains that shape modern taxonomy (Woese et al. 1990).

Due to the vast array of phylogenetic applications that depend on phylogenetic trees, it is imperative to understand and minimize sources of phylogenetic error to ensure accurate interpretations. The previous chapter illustrates the effect of methodological choices on 16S phylogenetic diversity and topological accuracy. We established a software framework to simulate short reads that would be obtained from high-throughput DNA sequencers by leveraging the high-quality SILVA LTP full-length 16S sequence database and conduct comparisons between 16S phylogenies reconstructed under different parameters. We examined the relationship between sequence length and phylogenetic error at hypervariable regions V2, V3, V4, V5, and V6, stepwise across each location in the alignment, with differing levels of phylogenetic diversity, tree size, and both with and without full-length reference sequences. Throughout these analyses, we demonstrated that the inclusion of reference sequences increases phylogenetic accuracy with remarkable improvements in the similarity of phylogenies reconstructed using short reads to phylogenies inferred from full-length gene sequences. Importantly, we measure the accuracy of phylogenetic tree reconstruction both in with phylogenetic diversity and tree topology metrics to ensure maximum applicability to researchers.

Taken together, these results demonstrate the effects of both the highly recognized and the understudied contributors to the accuracy of phylogenetic tree reconstruction in microbiome studies. We underscore expense-free improvements to phylogenetic accuracy through incorporating reference sequences which are particularly beneficial to short read studies. Finally, we researchers should consider how these methodological choices impact phylogenetic accuracy and subsequent phylogeny-driven inferences.

References

1. Clarridge JE. 2004. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews.
2. Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. 2018. Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota. mBio.
3. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.
4. Hugenholtz P, Goebel BM, Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774.
5. Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology.
6. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology.
7. Thompson LR, The Earth Microbiome Project Consortium, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature.
8. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ.
9. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences.
10. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research.