ABSTRACT

Title of dissertation: Assessing 16S rRNA Marker-Gene Survey
Measurement Process Using
Mixtures of Environmental Samples

Nathanael D. Olson
Doctor of Philosophy, 2018

Dissertation directed by: Professor Héctor Corrada Bravo
Department of Computer Science

Microbial communities play a fundamental role in environmental and human health. Targeted sequencing of the 16S rRNA gene, 16S rRNA marker-gene surveys, is used to measure and thus characterize these communities. The 16S rRNA marker-gene survey measurement process includes a number of molecular laboratory and computational steps. A rigorous measurement assessment framework can evaluate measurement method performance, in turn improving the validity of marker-gene survey study conclusions. In this dissertation, I present a novel framework and mixture dataset for assessing 16S rRNA marker-gene survey bioinformatic methods. Additionally, I developed software to facilitate working with 16S rRNA reference sequence databases and 16S rRNA marker-gene survey feature data. Computational steps, collectively referred to as bioinformatic pipelines, combine multiple algorithms to convert raw sequence data into a count table, which is subsequently used to test biological hypotheses. Algorithm choice and parameters can significantly impact pipeline results. The assessment framework and software developed for this dis-

sertation improve upon existing assessment methods and can be used to evaluate new computational methods and optimize existing pipelines. Furthermore, the assessment framework presented here can be applied to other microbial community measurement methods such as shotgun metagenomics.

Assessing 16S rRNA Marker-Gene Survey Measurement Process Using Mixtures of Environmental Samples

by

Nathanael D. Olson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Héctor Corrada Bravo, Chair/Advisor
Professor Mihai Pop
Dr. Marc L. Salit
Professor O. Colin Stine
Professor Nathan Swenson, Dean's Representative

Acknowledgments

This work was funded by the National Institute of Standards and Technology and an NIH R01HG005220 grant awarded to Dr. Héctor Corrada Bravo. I want to thank my committee, Héctor Corrada Bravo, Mihai Pop, Marc Salit, O. Colin Stine, and Nathan Swenson.

My committee included experts in computer science, applied mathematics, microbiology, community ecology, and metrology. Their diverse areas of expertise helped make this work possible. I especially want to thank my advisor, Héctor, for teaching me to always ask the next question, dig deeper into the data, and also for encouaging me to push the project out into the world. Thank you to Dr. Mihai Pop for teaching a biologist string and network algorithms. Thanks to Dr. Misaki Takabayshi, my undergraduate research advisor, for introducing me to research and the dark art of PCR. Thanks to Dr. Jayne Morrow for encouraging me to go back to school and for constant support. The Corrada-Bravo and Pop labs are filled with brilliant and kind people, with whom it was a pleasure and honor to work.

This would would not have been possible without the help of many collaborators: Dr. Joe Paulson, Justin Wagner, Nidhi Shah, Jayaram Kancherla, and Daniel Hwang helped develop metagenomeFeatures. David Catoe, Shan Li, Stephanie Hao, and Dr. Winston Timp helped to generate the assessment dataset. M. Senthil Kumar and Jacquelyn S Meisel worked on the abundance and beta-diversity assessment chapters respectively. I also want to thank Dr. Prachi Kulkarni for editing this dissertation.

I want to thank my parents for encouraging curiosity, letting me find my own way through life, and for providing me with love and music. I also want to thank my Grandfather, who is a role model and a constant source of inspiration.

Most of all I want to thank my wife, Erica, and son Charlie, for loving me through the hard times. Thanks to Charlie for letting me read microbiome papers to you and for listening to me as I worked though roadblocks in my research. More importantly, for having an infectious laugh that made it easy to take a break and forget about work for a little while.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Microorganisms, the unseen majority, play an important role in environmental and human health. Globally there are an estimated 4 to 6×10^{30} prokaryotic cells on earth driving processes such as the carbon and nitrogen cycle [104, 46, 104]. In the human body, bacterial cells are as abundant as human cells [90] and aid in fundamental processes such as digestion [23, 111]. The human gut microbiome has been linked to numerous diseases ranging from inflammatory bowel disease to autism [23, 94]. Biotherapeutics, probiotics developed to treat disease, have the potential to revolutionize medicine and treat microbiome-linked diseases [111]. For example, fecal microbiome transplants have been successfully used to treat recalcitrant *Clostridium dificile* infections [81]. Accurately characterizating microbial communities is critical to successful biotherapeutics development.

Recent advances in deoxyribonucleic acid (DNA) sequencing technology has changed how we measure and thus study, microbiomes. The two most commonly used methods to characterize microbial communities are shotgun metagenomics and 16S ribosomal ribonucleic acid (rRNA) marker-gene surveys [52]. Shotgun metagenomics is the random sequencing of all genomic material in a sample. 16S rRNA marker-gene surveys use targeted sequencing to characterize community taxonomic composition. There are benefits and disadvantage to both methods [25]. Shotgun sequencing is useful for functional information and strain-level analysis, whereas marker-gene sequencing provides a more complete view of the community composition. Additionally, shotgun sequencing is significantly more expensive than marker-gene surveys and is, therefore, cost prohibitive for large cohort and global studies, such as the human microbiome project [99] and earth microbiome project [98]. The focus of this dissertation is assessing 16S rRNA marker-gene survey methods. However, the approaches presented apply to shotgun metagenomic method assessment as well.

1.1 16S rRNA Marker-Gene Surveys

Marker-gene survey data is used to characterize both organismal and community level differences [34]. Differential abundance testing is a type of organismal level analysis used to identify organisms associated with specific treatment conditions, for example, a pathogen associated with a disease state. Ecological diversity metrics are used to characterize microbial community richness and evenness within individual samples (alpha-diversity) and sample pairwise similarity (beta-diversity).

Marker-gene survey data is collected through a complex multi-step measurement process [34]. The measurement process consists of numerous laboratory and computational steps. Laboratory steps include DNA extraction, polymerase chain reaction (PCR), library preparation, and sequencing. Computational steps include pre-processing, feature inference, feature annotation, and normalization. See Sections 2.1 and 2.2 for a detailed description of the measurement processes. Limitations in our understanding of the measurement process impede microbiome research.

1.2 Measurement Assessment

Measurement assessment is used to characterize and validate the measurement process and as a result, advance microbiome research. Results from marker-gene surveys indicate a potential connection between obesity and the gut microbiome [54]. However, a meta-analysis combining data from multiple studies only found a weak association between obesity and the gut microbiome [95]. With a well-characterized measurement process, we can better evaluate study conclusions. In turn, reducing the identification of spurious associations, such as the ones identified by the obesity studies analyzed in the Sze and Schloss [95] meta-analysis.

A key component of measurement assessment is data with an expected value. Common sources of data used to assess marker-gene surveys include computer simulated data (*in silco*), mixtures of DNA or cells from individual organisms (mock communities), and technical replicates of environmental samples. *In silico* data and mock communities provide expected values for use in assessment but do not recapitulate the complexity of environmental samples or the error profile of real sequencing data. Without an expected value, technical replicates are only suitable for method comparison. Another data type, mixtures of environmental samples, provide the complexity of real data and an expected value for assessment.

1.3 Dissertation Summary

In this dissertation, I will present a framework for 16S rRNA marker-gene survey measurement assessment using a novel mixture dataset along with software to facilitate working with 16S rRNA marker-gene survey data. Chapter 2 provides an overview of the 16S rRNA sequencing measurement process and assessment. In Chapter 3, I describe the development and application of an assessment framework for evaluate the relative and differential abundance values for count tables generated using different bioinformatic pipelines. For Chapter 4, I developed methods to assess beta-diversity. I then used the methods to evaluate the impact of sequencing artifacts on bioinformatic pipelines and normalization methods. Chapters 3 and 4 utilize a two-sample titration assessment dataset generated are part of this disseration (Fig. 1.1). Finally, Chapter 5 describes the Bioconductor R package, *metagenomeFeatures*, I developed for working with 16S rRNA marker-gene survey data and 16S rRNA databases. The assessment framework and software along with the assessment study results presented here will facilitate the development of improved computational methods and advance our understanding of the marker-gene survey measurement process. The last three sections of this chapter provide a brief overview of the three main disseration chapters.

1.3.1 Abundance Assessment

The proper measurement method evaluation characterizes the impact of individual steps on the measurement process. Furthermore, it also helps identify where

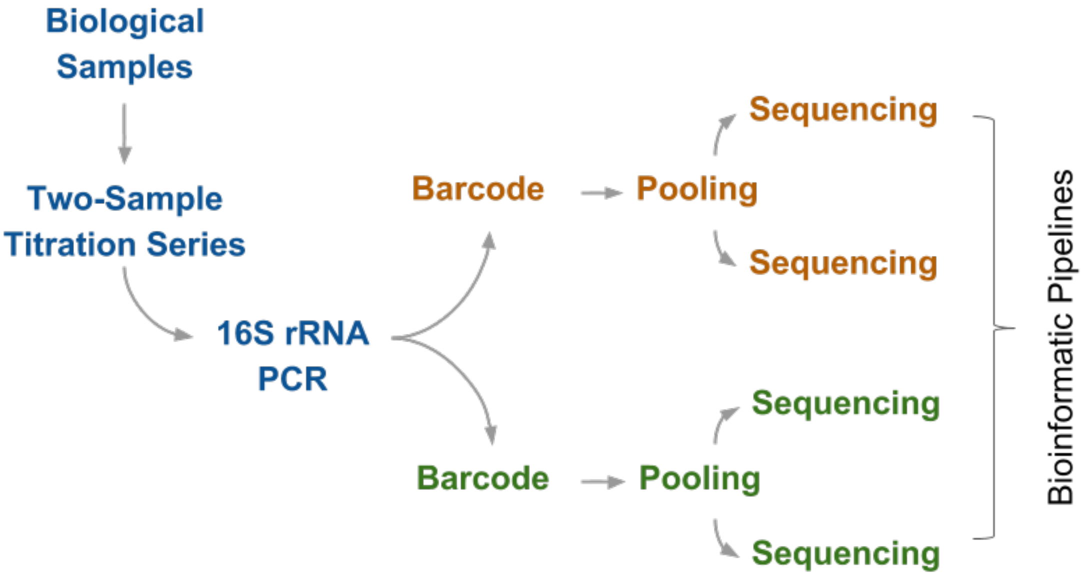


Figure 1.1: Assessment dataset experimental design flowchart. Two-sample titration series were generated using DNA from stool samples collected as part of an Enterotoxigenic *Escherichia coli* (ETEC) vaccine trial. The titration end point samples were selected as the initial sampling timepoint in the vaccine trial and timepoint after exposure with the highest measured concentration of ETEC. The assessment dataset includes multiple biological factors and technical replicates enabling the charaterization of multiple sources of bias and dispersion in the measurement process.The assessment dataset includes samples from five subjects, vaccine trial participants, with a two-sample titration series for each subject. Four 16S rRNA PCR replicates per titration series sample. The PCR assays were split into technical replicates and sent to two laboratories for library preparation and sequencing. Libraries were sequenced twice as each laboratory for a total of four sequence datasets.

to focus efforts for improving the measurement process. Appropriate datasets and methods are needed to evaluate the 16S rRNA marker-gene-survey measurement process. Numerous studies have qualitatively and quantitatively assessed the 16S rRNA measurement process using mock communities, simulated data, and environmental samples.

Qualitative characterisitcs are commonly assessed using mock communities [10]. As the number of organisms in the mock community is known, the total number of features can be compared to the expected value. The number of observed features in a mock community is often significantly higher than the expected number of organisms [50]. Although, when mock community data are processed using sequence inference methods the count tables, the observed and expected number of features is consistent [19]. The higher than expected number of features is often attributed to sequencing and PCR artifacts as well as reagent contaminants [14, 42]. However, benchmarking studies evaluating sequence inference methods, such as DADA2, aim to reduce the number of features due to sequence artifacts. While mock communities have a known value, they lack the feature diversity and relative abundance dynamic range of real samples [10].

The quantitative characteristics of 16S rRNA sequence data are normally assessed using mock communities and simulated data. To assess the quantitative accuracy of relative abundance estimates, mock communities of equimolar and staggered concentration are commonly used [50]. Results from relative abundance estimates using mock communities generated from mixtures of DNA have shown taxonomic specific effects where individual taxa are under or over represented in a sample.

These taxonomic specific effects have been attributed to primer mismatches and DNA extraction biases [14]. Simulated count tables have been used to assess differential abundance methods, where specific taxa are artificially overrepresented in one set of samples compared to another [62]. Only computational steps of the measurement process can be assessed using simulated data.

Quantitative and qualitative assessment can also be performed using sequence data generated from mixtures of environmental samples. Using simulated data and mock communities, evaluating and benchmarking new methods can result in over training the bioinformatic pipelines to data that do not recapitulate the sequencing error profile and feature diversity of real samples. Data from environmental samples, which are real samples, are often used to benchmark new molecular laboratory and computational methods. However, without an expected value with which to compare, only measurement precision can be evaluated. By mixing environmental samples, an expected value can be calculated using information from the unmixed samples and the mixture design. Mixtures of environmental samples have previously been used to evaluate gene expression measurements, e.g. microarrays and RNAseq data [71, 76, 97].

In Chapter 3, we describe the mixture dataset of extracted DNA from human stool samples we generated for assessing 16S rRNA sequencing. The mixture datasets were processed using three bioinformatic pipelines. Using the mixture dataset, we developed novel methods to assess the qualitative and quantitative characteristics of the pipeline results. The quantitative results were similar across pipelines but the qualitative results varied by pipelines.

1.3.2 Diversity Assessment

Biases introduced during the marker-gene survey measurement process can impact downstream analyses such as beta-diversity. Bioinformatic pipelines and normalization methods are often used to reduce biases introduced during the marker-gene survey measurement process [34, 49].

Bioinformatic pipelines reduce bias by removing sequencing artifacts, such as single and multi-base pair variants, and chimeric sequences, from microbiome datasets. If not accounted for, these artifacts may incorrectly be attributed to novel diversity in a sample. Bioinformatic pipelines use clustering or sequence inference techniques to group reads into biologically informative units. Standard clustering methods include *de novo* clustering of sequences based on pairwise sequence similarities [88] and closed-reference clustering of reads against a reference database [29]. Open-reference clustering is a combination of the two, first applying a closed-reference approach, followed by *de novo* clustering of reads that did not map to a reference [82]. Sequence inference methods use statistical models and algorithms to group sequences independent of similarity, but based on the probability that a lower abundant sequence is an artifact originating from a more highly abundant sequence [19, 1]. The resulting features, operational taxonomic units (OTUs) for clustering methods, and sequence variants (SVs) for sequence inference methods, have different characteristics because the methods vary in their ability to detect and remove errors while retaining true biological sequences.

Rarefaction and numeric normalization methods account for differences in sample total abundance caused by uneven sample pooling prior to sequencing and differences in sequencing run throughput. Rarifying abundance data traces its origins to macroecology, where counts for a unit (sample) are randomly subsampled to a user-defined constant level [35]. Although there are concerns about its statistical validity [62], rarefaction is currently the only normalization method for unweighted, presence-absence based, beta-diversity metrics [102]. For weighted, abundance based, beta-diversity analyses, we can apply numeric normalization methods, such as total and cumulative sum scaling (TSS and CSS), where counts are divided by sample total abundance (TSS) or by the cumulative abundance (CSS) for a defined percentile [72]. CSS is one of the few normalization methods developed specifically for 16S rRNA marker-gene survey data. Other normalization methods, including upper quartile (UQ), trimmed mean of M values (TMM) and relative log expression [84, 58], were initially developed for normalizing RNAseq and microarray data. Many studies have found these methods useful in normalizing marker-gene survey data for differential abundance analysis, though it is unclear whether these techniques are also suitable for beta-diversity analysis.

Beta-diversity is calculated using a variety of metrics that can be grouped based on whether they account for phylogenetic distance or not and feature relative abundance or presence/absence. The UniFrac metric was developed for marker-gene survey data and incorporates phylogenetic relatedness by comparing the branch lengths of features that are unique to two communities [37]. Unweighted UniFrac uses presence-absence information, whereas weighted UniFrac incorporates feature relative abundance. Taxonomic metrics do not consider the relationship between features. The Bray-Curtis and Jaccard dissimilarity indices are examples of weighted and unweighted taxonomic metrics respectively, as they do not consider the phylogenetic relationship between features [13, 44]. These four groups of beta-diversity metrics measure different community characteristics. The metrics should therefore be evaluated in a complementary manner to gain maximal insight into community-level sample differences [2].

Previous studies have evaluated different bioinformatics pipelines [92] and normalization methods [62, 102] using beta-diversity metrics. Yet, how well pipelines account for low sequence quality and total abundance differences remains unknown. The mixture dataset includes multiple levels of technical replication, allowing us to evaluate (1) beta-diversity repeatability, (2) ability to distinguish between groups of samples with varying similarity, and (3) identify differences in beta-diversity between biological and technical factors. Furthermore, the dataset includes data from four sequencing runs with different sequencing error rates and library sizes, enabling assessment of how each pipeline and method performs on varying quality datasets.

1.3.3 *metagenomeFeatures*

A key step in 16S rRNA marker-gene survey measurement process is comparing representative sequences to a reference database for taxonomic classification or phylogenetic placement [65]. There are numerous 16S rRNA reference databases, of which Greengenes, Ribosomal Database Project (RDP), and SILVA are arguably the most commonly used [28, 26, 79, 59]. Additionally, there are smaller system-specific databases such as the Human Oral Microbiome Database (HOMD) for the human oral microbiome [22, http://www.homd.org/] and soil reference database [24]. System-specific databases can improve taxonomic assignments for microbial communities not well represented in the major databases [85].

16S rRNA databases differ in sequence number and diversity, the taxonomic classification system, and the inclusion of intermediate ranks [5]. Databases format their data differently and use sequence identification systems unique to their database, challenging membership and composition comparisons. For example, Yang, Wang, and Qian [110] used the SILVA database to evaluate impact of different 16S rRNA variable regions on phylogenetic analysis. Similarly, Martinez-Porchas et al. [57] evaluated sequence similarity between 16S rRNA gene conserved regions also using the SILVA database. Differences in database formatting present a significant barrier to performing the same analysis using multiple databases. Additionally, taxonomic assignments can be database-dependent, providing further justification for database comparisons [75]. To facilitate database comparisons RNACentral a resource combining non-coding RNA databases provides a set of cross database sequence identifiers [96, http://rnacentral.org/].

We developed the Bioconductor R package *metagenomeFeatures* for working with both 16S rRNA gene databases and marker-gene survey feature data. *metagenomeFeatures* provides a common data structure for working with 16S rRNA databases and marker-gene survey feature data. Additionally, this package is the first step towards the development of a common data structure for use in analyzing metagenomic and marker-gene survey data using R Bioconductor packages such as *phyloseq* [60] and *metagenomeSeq* [73].

CHAPTER 2

Measurement Process and Assessment.

Metrology, measurement science, provides a well established framework for validating, characterizing and evaluating measurements [55, 33]. This framework was established for routine single measurand assays with well defined applications. Measurement validation is used to determine if a measurement process meets a set of specific requirement. The measurement process requirements can include measurement bias and precision are within a defined range or a minimum detectable analyte concentration. For example, measurement validation is used to determine if an assay is able to detect a specific pathogen in a stool sample at a minimum concentration. Measurement validation also requires that the sample used to validate the measurement is representative of the sample type the measurement process will encounter. The samples used for validation must also have known charateristics for use in evaluating measurement results. As part of the measurement validation process measurement expanded uncertainty is quantified determined. Measurement expanded uncertainty is obtained by first defining the measurand, decomposing the measurement process, and identifying sources of uncertainty [33]. Each source of uncertainty is quantified by experimental estimation, modelling from theoretical principles, or estimation using expert judgement. The expanded uncertainty is calculated by combining uncertainty estimates for individual sources. The observed measurement value and its expanded uncertainty along with the measurement requirements are used to validate the measurement process.

With 16S rRNA sequencing measurements are made for hundereds to thousands of organisms simultaneously. Furthermore, there is no clear single application as the same data are used to test multiple hypothesis. The traditional framework for characterizing and evaluating a measurement process is not easily applied. Therefore, I present a more general characterization of the measurement method performance as measurement assessment, which borrows from the measurement validation process. Where a measurement process is assessed relative to the measurand, the entity quantified by the measurement process [9]. Measurement process assessment consists of three steps; decomposing the measurement process, designing experiments to isolate measurement process elements, and evaluating bias and dispersion at each element. In this chapter I will describe and decompose the 16S rRNA sequencing measurement process. After decomposing the measurement process I will describe common 16S rRNA measurands including feature abundance, relative abundance, as well as alpha- and beta-diversity. Next, I will describe steps for conducting an assessment experment to characterize measurement process elements. Finally, I will place the assessment work presented in this dissertation within the larger context of 16S rRNA measurement process assessment.

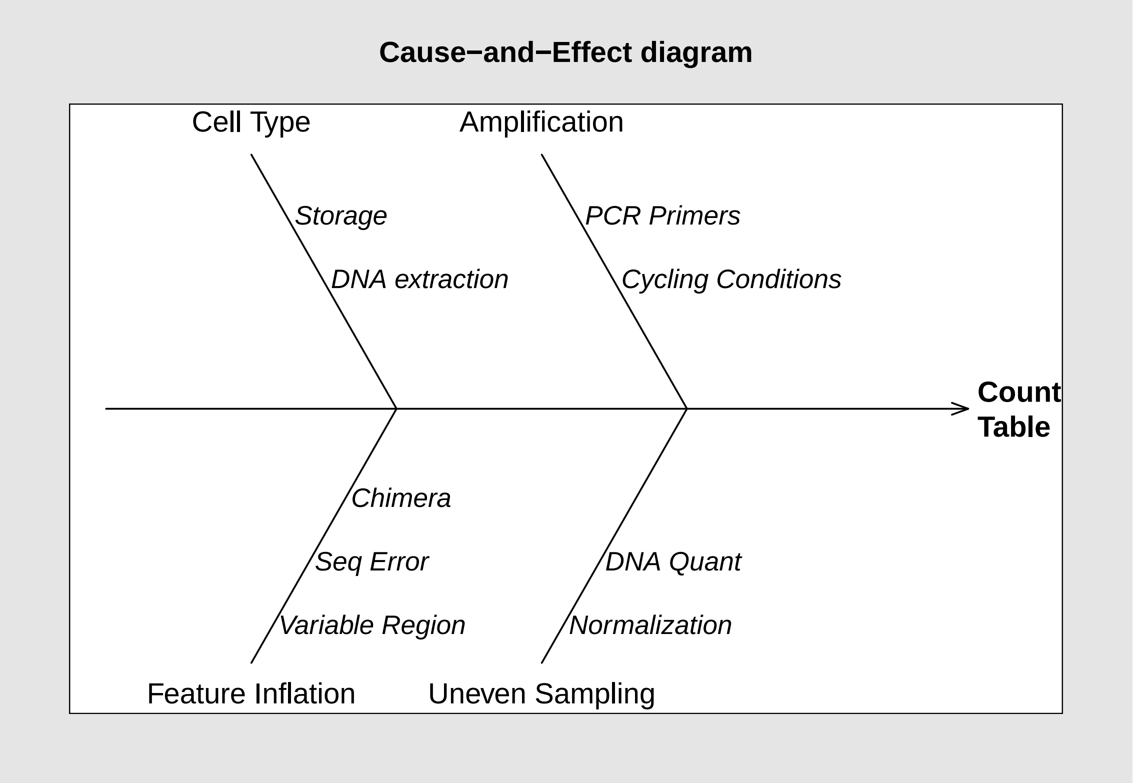


Figure 2.1: Cause-effect diagram for the 16S rRNA sequencing measurement process. Causes of bias and variability are indicated as branches, e.g., cell type, and different sources of error are indicated by branch labels, e.g., storage conditions.

2.1 Measurement Process

The 16S rRNA marker-gene survey measurement process includes a number of laboratory and computational steps (Fig. 2.2 and 2.3). In this section, I will provide an overview of the measurement process, and highlight significant known sources of error and dispersion (Fig. 2.1). See Goodrich et al. [34] for a review of 16S rRNA marker-gene survey measurement process and general recommendations for conducting microbiome experiments.

The 16S rRNA gene is used for marker-gene surveys as it is found in all prokaryotic organisms, including both bacteria and archaea and contains hypervariable and conserved regions. The conserved regions allow for the use of PCR primers which can amplify the 16S rRNA gene from diverse taxonomic groups [57, 48]. Whereas hypervariable regions, allow genus and sometime species-level taxonomic classification. Additionally, as the 16S rRNA gene is well studied there are several extensive well-curated databases, e.g. SILVA, RDP, and Greengenes [79, 59, 26]. There are drawbacks to using the 16S rRNA gene as well. Due to the conserved nature of the gene, sequences cannot be used for the strain-level taxonomic classification required for some applications, such as pathogen detection and identification [45]. Additionally, 16S rRNA is a multicopy gene that is known to be horizontally transferred between organisms, as a result, the within-genome 16S rRNA gene copy diversity can be greater than the between-genome diversity for some genera [74].

For the measurement process, laboratory component raw sequence data are generated from environmental samples (Fig. 2.2). Samples are initially collected and stored to minimize changes to the community composition prior to sample processing. Next, DNA is extracted from the sample, and the 16S rRNA is amplified using PCR. Then PCR products concentrations are normalized, diluted to a standard concentration, to minimize between sample variability in the number of reads obtained per sample. Finally, the normalized PCR products are pooled and sequenced. Preferential DNA extraction and PCR amplification are two of the largest sources of bias in the measurement process (Fig. 2.1). Other sources of error introduced at this point include PCR artifacts such as amplification errors (point mutations) and chimeras, sequence contaminants, and sequencing errors. Computational methods correct for PCR artifacts, contaminant removal, and sequencing

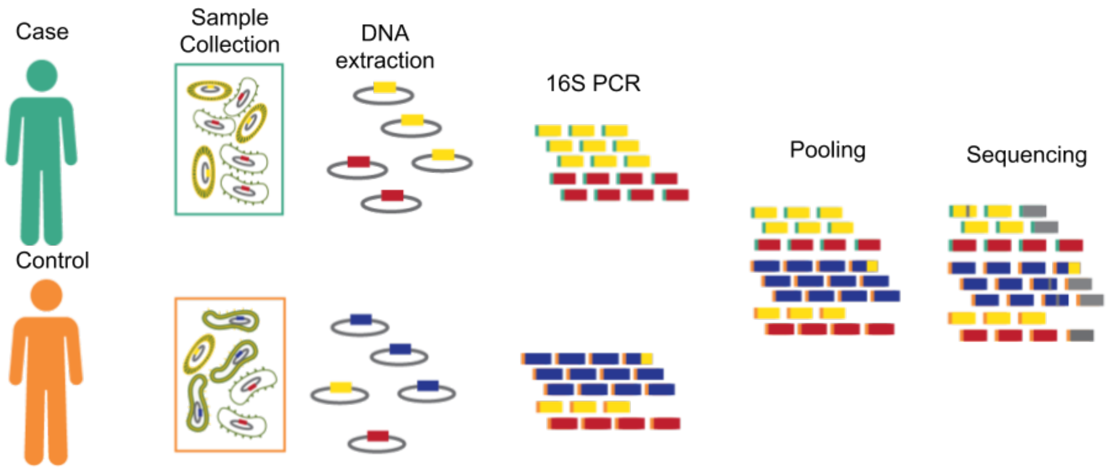


Figure 2.2: Diagram of the 16S rRNA marker gene survey measurement process for a simplified example study comparing case and control treatment groups. Samples are first collected from study participants. Colored boxes contains the true microbial community composition indicated by cartoon cells. The different cell walls and 16S rRNA sequence (grey oval with colored bar) indicate distinct biologically relevant units. Next, DNA is isolated from the other cellular material. Extraction efficiency differences are indicated by differences in the number of red, yellow, and blue bars relative to the sample numbers. PCR is used to target the 16S rRNA gene. Amplification efficiency biases shown as differences in the proportion of red, yellow, and blue PCR products from the DNA extract. Unique sample barcodes added during PCR are indicated as teal and orange bars on PCR amplicons. Chimeras, a PCR artifact, are shown as half blue and yellow PCR products. Sequencing libraries are produced by *Pooling* PCR products from the two samples. Uneven pooling can result in the under-representation of a sample (teal in this example) in the sequence dataset. The resulting libraries are sequenced, sequencing artifacts are indicated as grey reads, for failed sequences, and grey vertical lines for single base sequencing errors.

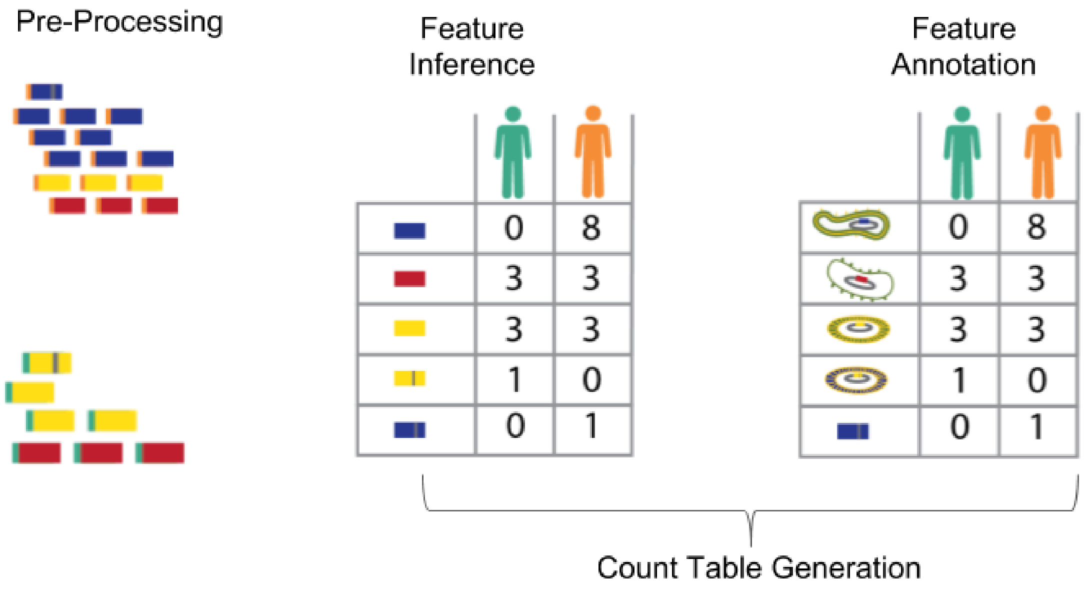


Figure 2.3: Diagram of the 16S rRNA marker gene survey measurement process computational steps. Pre-processing assigns sequencing reads to samples using the unique barcodes and removes low-quality reads as well as filters chimeras. Feature inference is used to group the pre-processed reads into biologically relevant units. Sequencing errors can result in spurious features if not assigned to the source feature. The yellow and blue sequences with a grey vertical line are spurious features. Finally, taxonomic assignment is performed as part of feature annotation. Spurious features can be assigned to the wrong organism, yellow bar with grey line, or be unassigned, blue bar with grey line.

error correction. Currently, there are no computational or laboratory methods to correct for preferential extraction or amplification.

Computational methods, collectively referred to as the bioinformatic pipeline, convert raw sequence data into an annotated count table for use in downstream analysis (Fig. 2.2). Bioinformatic pipelines use the same general approach, though the methods and order of individual steps vary by pipeline. The first step is pre-processing the raw reads. Pre-processing includes initial quality control steps, preparing reads for feature inference. Next, feature inference is performed, grouping pre-processed into biologically relevant units. Finally, feature annotation is used to obtain information about feature taxonomy and phylogenetic relatedness. Biases introduced by the computational methods are either due to failure to correct for biases from the laboratory component or errors in feature annotation.

The annotated count table is then used in feature-level and community-level downstream analysis (Fig. 2.4). Differential abundance is the most commonly used feature-level analysis. Differential abundance is used to estimate feature relative abundance between treatments [72]. Alpha- and beta-diversity analyses are the most commonly used type of community analyses. Alpha-diversity metrics are numeric summaries of individual samples, including richness, evenness, and phylogenetic diversity metrics. Beta-diversity is a measure of sample similarity and is used to compare feature presence-absence or incorporate relative abundance and phylogenetic relatedness.

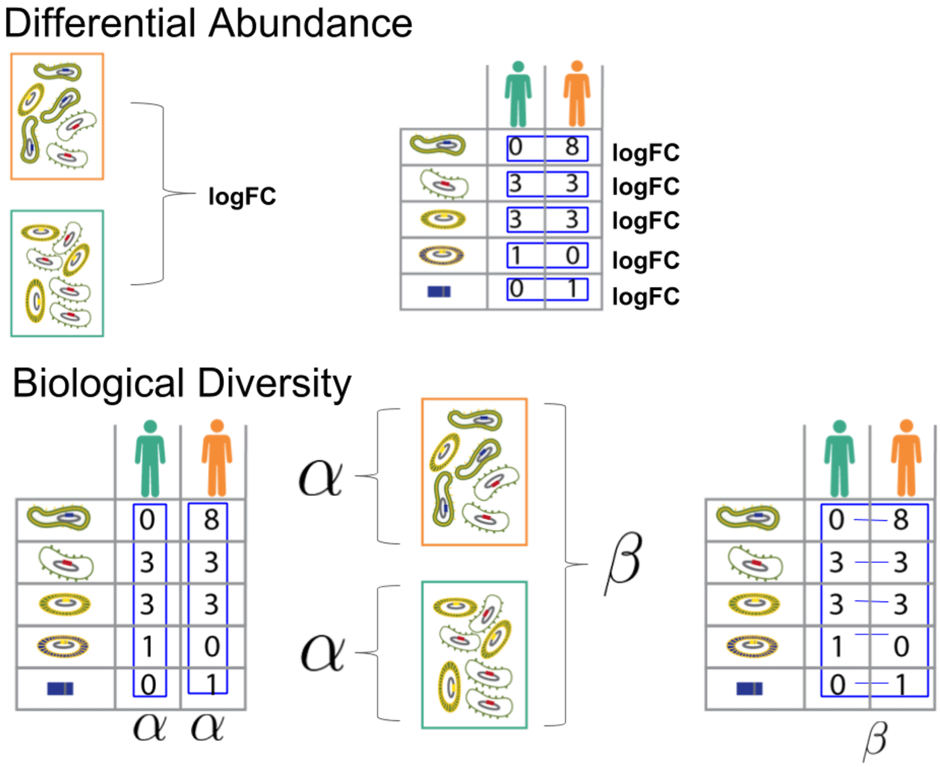


Figure 2.4: Diagram of 16S rRNA marker gene survey differential abundance and biological diversity downstream analysis methods. Differential abundance is a row-wise operation comparing feature abundance between treatment conditions with log fold-change (logFC) estimates calculated for each feature. Alpha (α) diversity is a column-wise operation where a single value is calculated per sample. Beta (β) diversity is a row- and column-wise operation, with a single value calculated for each sample pair, using paired feature information.

2.2 Feature Inference and Normalization

The following section provides additional details about the feature inference and computational normalization methods that are the focus of the assessment framework presented in this disseration. Feature inference is used to group sequences into features that are representative of biologically relevant units. Computational normalization is used to reduce biases in downstream analysis due to differences in total abundance between samples. The feature inference process should maximize the number of features representing true sequences while minimizing the number of features representing sequence artifacts. The two primary feature inference method types are distance-based clustering and sequence inference.

Distance-based feature inference methods cluster sequences based on defined similarity thresholds. These clustering thresholds are based on taxonomic group sequence diversity and sequencing error (e.g., 99% species level, 97% genus level, 1% error rate). Though studies characterizing within species and genera 16S rRNA sequence diversity [74], and sequencing error rates [42, 86, 27] challenge the validity of these values [31].

There are three approaches to distance-based clustering, *de novo*, closed-reference, and open-reference. For *de novo* clustering, the pre-processed sequences are clustered based on the desired threshold. For closed-reference clustering, the sequences are assigned to reference sequences based on the defined similarity threshold. The reference sequences are sequences from a reference database previously clustered at the desired threshold. Open-reference clustering combines *de novo* and closed-reference clustering. Reads are first clustered using closed-reference clustering then reads not assigned to reference clusters are clustered *de novo*. A limitation of distance-based feature inference methods is the use of a defined similarity threshold. Sequence inference methods were developed to avoid having to define such thresholds.

Sequence inference methods, another type of features inference, use statistical models or heuristic algorithms to infer the true biological sequence from which a read was generated. By inferring the true biological sequence, this method avoids having to define an arbitrary threshold for grouping sequences. DADA2, a sequence inference method, uses a probabilistic model and an expectation-maximization algorithm to test whether less abundant sequences are derived from higher abundance sequences [19]. DADA2 is the only sequence inference method to consider base quality scores. DADA2 is computationally expensive on a per sample basis, though scalable as individual samples can be run in parallel. Other sequence inference methods use heuristics, reducing the method the computational cost. Similar to DADA2, UNOISE2 uses a model to assign lower abundance sequences to higher abundance sequences [30]. However, UNOISE2 uses a single function with parameter values set using training data. UNOISE2 significantly reduces the computational cost by using model parameters defined *a priori*, eliminating the need for the expectation-maximization step. Deblur proportionally assigns lower abundance sequences to higher abundance sequences using estimates for the number of lower abundance sequences that are error derived sequences [1].

After feature inference, the resulting count tables are normalized prior to use in downstream analysis. Count table mormalization methods minimize biases due to differences in the total feature abundance between samples. Variability in total feature abundance is due to differences in the number of reads generated per sample, and the proportion of sequences passing quality filtering. There are two types of normalization methods rarefaction and scaling. Rarefaction has its roots in ecology, where it is used to normalize sampling intensity per survey area [35, 40]. To rarify counts, individual sample counts are randomly subsampled, without replacement, to a user-defined level and samples with total abundance values less than the rarefaction level are dropped. Scaling methods normalize count table values by sample-specific normalization factors. For total sum scaling (TSS), and cumulative sum scaling (CSS), sample counts are divided by the sum of feature counts to a defined abundance percentile, 0.75 for CSS, and total abundance for TSS (proportions) [72]. Other commonly used normalization methods include relative log expression (RLE), trimmed mean of M values (TMM), and upper quartile normalization [84, 58]. These methods were developed for normalizing microarray and RNAseq data but have been successfully used to normalize 16S rRNA marker-gene survey data [62].

2.3 Conducting an Assessment Experiment

Measurement assessment experiment are used to characterize the bias and dispersion for elements of a measurement process. There are four steps to designing a measurement assessment experiment; (1) define the measurand, (2) identify appropriate data and metrics for the assessment, (3) design an experiment to isolating element(s) of a measurement process for assessment.

First define the measurand when developing an assessment method. The measurand is the observed value being quantified [9]. For marker-gene survey analysis, there are multiple measurand definitions. The measurand can be defined based on the downstream analysis, e.g., log fold-change for differential abundance analysis or diversity metric for alpha- and beta-diversity analysis. The measurands can also be more general, such as count table abundance values or feature DNA sequence.

After defining the measurand appropriate data and metrics are identified. The primary data types used in marker-gene survey measurement assessment include *in-silico* data, mock communities, and technical replicates. Different parts of the measurement process and measurands can be assessed with different data types. *in-silico* data, either simulated sequence data or count tables, can be used to assess computational methods and downstream analysis. Mock communities, either mixtures of DNA or cells, can be used to assess laboratory procedures, as well as computational methods. Technical replicates are used to assess laboratory and computational method repeatability. Once the measurand is defined, and appropriate data for assessment is identified, assessment is performed by comparing expected values to observed values.

Metrics used for assessment are dependent on the statistical properties of the measurand. For quantitative measurands, such as count table values, relative error (the difference between the expected and observed values, normalized by the expected value) is commonly used. For evaluating the linear relationship between observed and expected values a linear model can be used to estimate the overall agreement, R^2, and whether the relationship is 1-to-1, model slope. Another metric for quantitative measurands, is signal-to-noise ratio, for example, with beta-diversity metrics, the ratio of beta-diversity between and within treatments or conditions. For qualitative measures, such as feature presence-absence, standard truth table metrics, such as sensitivity and specificity can be used. Similarly, area under the curve (AUC) and receiver operating characteristics (ROC) curves can be used to evaluate qualitative performance for a set of observed values.

The next step is to determine measurement process steps or element one is interested in evaluating and design an experiment isolating the element. Assessments can compare the performance of different methods used in individual steps of the measurement process, for example, DNA extraction or PCR. Assessment methods can also evalute part of the measurement process, such as the computational or laboratory components. Alternatively, one can assess the measurement process as a whole. The part of the measurement process being evaluated defines the experimental design used to generate the assessment dataset.

2.3.1 Assessment using Mixtures of Environmental Samples

Limitations to using environmental samples and mock communities for measurement assessment can be addressed using mixtures of environmental samples. Mixtures of environmental samples provide the complexity of real data regarding feature diversity and dynamic range with expected values for comparison. Mixtures of environmental samples have previously been used to evaluate microarray and RNAseq methods [43, 97, 76, 71]. When used to assess marker-gene surveys, mixtures of environmental samples can be used to assess count table accuracy, differential abundance, and beta-diversity. Simulated mixtures of environmental samples (simulated count tables) have previously been used to evaluate deferentially abundant features and beta diversity [62].

A limitation to using mixtures of environmental samples is the uncertainty in the expected values. Expected value estimates are based on information from the unmixed samples and the mixture design. Measurement values obtained for the unmixed samples are generally obtained using the same measurement process being assessed, and therefore, measurement uncertainty may not be well understood. Using the same measurement process and technical replicates allows for the quantification of the measurement uncertainty, but provides no information about potential measurement biases. In addition to uncertainty in the unmixed sample values, there may be uncertainty in the mixture design, both of which can be estimated. For RNAseq studies, since the sequencing assay targets mRNA, the proportion of mRNA in the RNA extract needs to be accounted for in the mixture equation [71]. Similarly for 16S rRNA marker-gene surveys, as the PCR assay targets prokaryotic DNA, the proportion of prokaryotic DNA in the sample should be characterized and taken into consideration.

CHAPTER 3

Abundance Assessment

Assessing 16S marker-gene survey data analysis methods using mixtures of human stool sample DNA extracts.

3.1 Abstract

16S rRNA marker-gene surveys use targeted sequencing to characterize prokaryotic microbial communities. Analysis of these studies is confronted with numerous bioinformatic pipelines and downstream analysis methods, with limited guidance on how to decide between appropriate methods from simulation studies or limited complexity benchmark studies. Appropriate data sets and statistics for assessing these methods are needed. A mixture of environmental samples is one approach for generating assessment data sets with the real data complexity while providing an expected value. We developed a mixture dataset for assessing 16S rRNA bioinformatic pipelines and downstream analysis methods using samples collected from participants in a Enterotoxigenic *Escherichia coli* (ETEC) vaccine trial participants. A two-sample titration mixture design was used where DNA from stool samples prior

to ETEC exposure was titrated into stools samples collected after exposure, in effect diluting the amount of ETEC in the mixed sample. The sequencing data were processed using multiple bioinformatic pipelines, DADA2 a sequence inference method, Mothur a *de novo* clustering method, and QIIME with open-reference clustering. The pipelines varied in the number of features and proportion of reads passing quality control but had similar sparsity. The mixture dataset was used to qualitatively and quantitatively assess the count tables generated using the pipelines. Statistical tests were used to determine if features only present in unmixed samples and titrations, *unmixed*- and *titration*-specific features, were had abundance value that could be explained by sampling alone. For Mothur and QIIME less than 5% of *unmixed*- and *titration*-specific feature abundance could not be explained by sampling alone where as for DADA2 greater than 50% of *unmixed*-specific features and 10% of *titration*- specific features could not be explained by sampling alone. The quantitative assessment evaluated pipeline performance by comparing observed to expected relative and differential abundance values. Expected relative abundance and differential abundance values were calculated using information from the unmixed samples and mixture design. Overall the observed relative abundance and differential abundance values were consistent with the expected values. We developed feature-level bias and variance metric to further characterize relative abudance and differential abundance quantitative performance. Relative abundance feature-level bias metric was significantly different across the three platforms with DADA2 having the lowest bias, followed by Mothur, and QIIME. The relative abundance feature-level variance metric and both the differential abundance feature-level bias and variance metrics did not differ significantly across the three pipelines. The dataset and methods developed for this study will serve as a valuable community resource for assessing 16S rRNA marker-gene survey bioinformatic methods.

3.2 Introduction

Targeted sequencing of the 16S rRNA gene, 16S rRNA marker-gene-surveys, is a commonly used method for characterizing microbial communities, microbiomes. The 16S rRNA marker-gene-survey measurement process includes molecular (e.g. PCR and sequencing) and computational steps (e.g., sequence clustering) [34]. Molecular steps are used to selectively target and sequence the 16S rRNA gene from prokaryotic organisms within a sample. The computational steps convert the raw sequence data into a matrix with feature (e.g., operational taxonomic units) relative abundance values for each sample [34]. Both molecular and computational measurement process steps contribute to the overall measurement bias and dispersion [27, 34, 15]. Proper measurement method evaluation allows for the characterization of how individual steps impact the measurement processes as a whole and determine where to focus efforts for improving the measurement process. Appropriate datasets and methods are needed to evaluate the 16S rRNA marker-gene-survey measurement process. A sample or dataset with "ground truth" is needed to characterize measurement process accuracy. Numerous studies have evaluated quantitative and qualitative characteristics of the 16S rRNA measurement process using mock communities, simulated data, and environmental samples.

To assess the 16S rRNA sequencing measurement process qualitative characteristics of a mock communities are commonly used [10]. As the number of organisms in the mock community is known, the total number of features can be compared to the expected value. The number of observed features in a mock community is often significantly higher than the expected number of organism [50]. The higher than expected number of features is often attributed to sequencing and PCR artifacts as well as reagent contaminants [14, 42]. A notable exception to this is mock community benchmarking studies evaluating sequencing inference method, such as DADA2 [19]. Sequence inference methods aim to reduce the number of sequence artifacts features. While mock communities have a known value, they lack the feature diversity and relative abundance dynamic range of real samples [10].

The quantitative characteristics of 16S rRNA sequence data are normally assessed using mock communities and simulated data. Mock communities of equimolar and staggered concentration are used to assess relative abundance estimate quantitative accuracy [50]. Results from relative abundance estimates using mock communities generated from mixtures of DNA have shown taxonomic specific effects where individual taxa are under or over represented in a sample. These taxonomic specific effects have been attributed to primer mismatches and DNA extraction biases [14]. Simulated count tables have been used to assess differential abundance method, where specific taxa are artificially overrepresented in one set of samples compared to another [62]. Using simulated data to assess log fold-change estimates only evaluates computational steps of the measurement process.

Quantitative and qualitative assessment can also be performed using sequence data generated from mixtures of environmental samples. While simulated data and mock communities are useful in evaluating and benchmarking new methods one needs to consider that methods optimized for mock communities and simulated data are not necessarily optimized for the sequencing error profile and feature diversity of real samples. Data from environmental samples, which are real samples, are often used to benchmark new molecular laboratory and computational methods. However, without an expected value to compare to, only measurement precision can be evaluated. By mixing environmental samples, an expected value can be calculated using information from the unmixed samples and how they were mixed. Mixtures of environmental samples have previously been used to evaluate gene expression measurements microarrays and RNAseq data[71, 76, 97].

In the present study, we developed a mixture dataset of extracted DNA from human stool samples for assessing 16S rRNA sequencing. The mixture datasets were processed using three bioinformatic pipelines. We developed metrics for qualitative and quantitative assessment of the bioinformatic pipeline results. The quantitative results were similar across pipelines but the qualitative results varied across pipelines. We have made both the dataset and metrics developed in this study publically available for evaluating new bioinformatic pipelines.

3.3 Methods

3.3.1 Two-Sample Titration Design

Samples collected at multiple timepoints during a Enterotoxigenic *E. coli* (ETEC) vaccine trial [38] were used to generate a two-sample titration dataset for assessing the 16S rRNA marker-gene survey measurement process. Samples from five trial participants were selected for our two-sample titration dataset. Trial participants (subjects) and sampling timepoints were selected based on *E. coli* abundance data collected using qPCR and 16S rRNA sequencing from Pop et al. [77]. Only individuals with no *E. coli* detected in samples collected from trial participants prior to ETEC exposure twere used for our two-samples titrations. Post ETEC exposure (POST) samples were identified as the timepoint after exposure to ETEC with the highest *E. coli* concentration for each subject (Fig. 3.1A). Due to limited sample availability, the timepoint with the second highest concentrations for E01JH0016 was used as the POST sample. Independent titration series were generated for each subject, where POST samples were titrated into PRE samples with POST proportions of 1/2, 1/4, 1/8, 1/16, 1/32, 1/1,024, and 1/32,768 (Fig. 3.1B). Unmixed (PRE and POST) sample DNA concentration was measured using NanoDrop ND-1000 (Thermo Fisher Scientific Inc. Waltham, MA USA). Unmixed samples were diluted to 12.5 $ng/\mu L$ in tris-EDTA buffer before making the two-sample titrations.

For our two-sample titration mixture design, the expected feature relative abundance can be calculated using equation (3.1), where θ_i, is the proportion of

POST DNA in titration i, q_{ij} is the relative abundance of feature j in titration i, and the relative abundance of feature j in the unmixed PRE and POST samples is $q_{pre,j}$ and $q_{post,j}$.

$$q_{ij} = \theta_i q_{post,j} + (1 - \theta_i) q_{pre,j} \tag{3.1}$$

3.3.2 Titration Validation

qPCR was used to validate volumetric mixing and check for differences in the proportion of prokaryotic DNA across titrations. To ensure that the two-sample titrations were volumetrically mixed according to the mixture design, independent ERCC plasmids were spiked into the unmixed PRE and POST samples [4] (NIST SRM SRM 2374) (Table 3.2). The ERCC plasmids were resuspended in 100 $ng/\mu L$ tris-EDTA buffer and 2 $ng/\mu L$ was spiked into the appropriate unmixed sample. Plasmids were spiked into unmixed samples after unmixed sample concentration was normalized to 12.5 $ng/\mu L$. POST sample ERCC plasmid abundance was quantified using TaqMan gene expression assays (FAM-MGB) (Catalog # 4448892, ThermoFisher) specific to each ERCC plasmid using the TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA). To check for differences in the proportion of bacterial DNA in the PRE and POST samples, titration bacterial DNA concentration was quantified using the Femto Bacterial DNA quantification kit (Zymo Research, Irvine CA). All samples were run in triplicate along with an in-house *E. coli* DNA log_{10} dilution standard curve. qPCR assays were performed

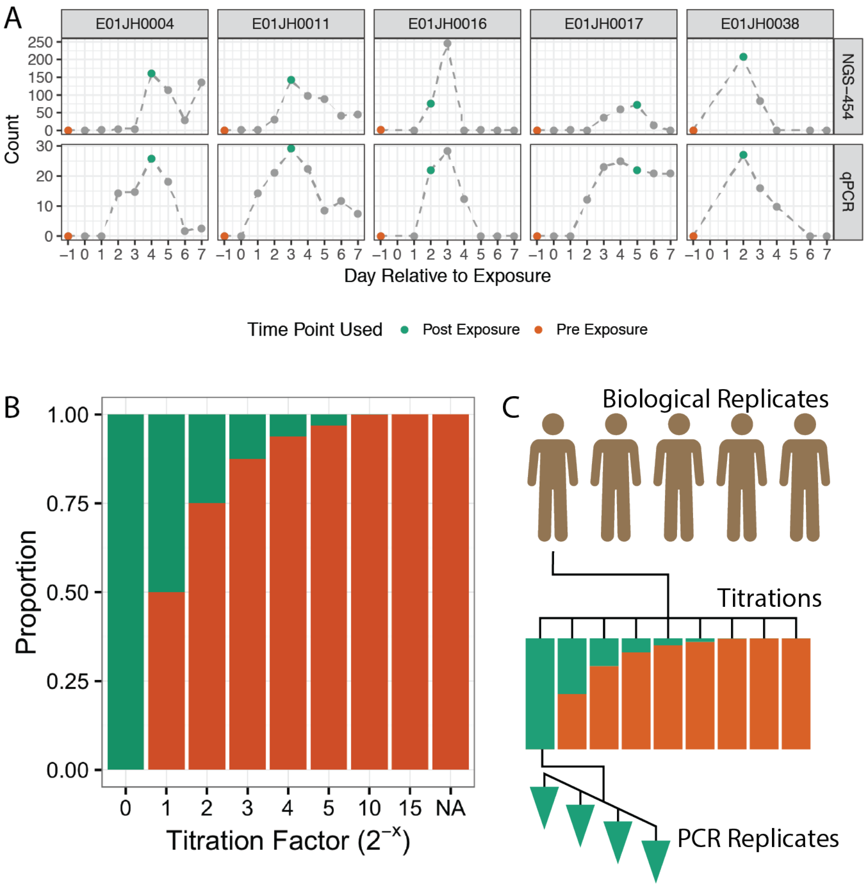


Figure 3.1: Sample selection and experimental design for the two-sample titration 16S rRNA marker-gene-survey assessment dataset. A) Pre- and post-exposure (PRE and POST) samples from five vaccine trial participants were selected based on *Escherichia coli* abundance measured using qPCR and 454 16S rRNA sequencing (454-NGS), data from Pop et al. [77]. PRE and POST samples are indicated with orange and green data points, respectively. Grey points are other samples from the vaccine trial time series. B) Proportion of DNA from PRE and POST samples in titration series samples. PRE samples were titrated into POST samples following a log_2 dilution series. The NA titration factor represents the unmixed PRE sample. C) PRE and POST samples from the five vaccine trial participants, subjects, were used to generate independent two-sample titration series. The result was a total of 45 samples, 7 titrations + 2 unmixed samples times 5 subjects. Four replicate PCRs were performed for each of the 45 samples resulting in 190 PCRs.

using the QuantStudio Real-Time qPCR (ThermoFisher). Amplification data and Ct values were exported as tsv files using QuantStudio Design and Analysis Software v1.4.1. Statistical analysis was performed on the exported data using custom scripts in R [80, https://github.com/nate-d-olson/mgtst_pub].

3.3.3 Sequencing

The 45 samples (seven titrations and two unmixed samples for each of five subjects) were processed using the Illumina 16S library protocol (16S Metagenomic Sequencing Library Preparation, posted date 11/27/2013, downloaded from https://support.illumina.com). This protocol specifies an initial 16S rRNA PCR followed by a sample indexing PCR, followed by normalization and sequencing.

A total of 192 16S rRNA PCR assays were run including four replicates per sample and 12 no-template controls, using Kapa HiFi HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA). The initial PCR assay targeted the V3-V5 region of the 16S rRNA gene, Bakt_341F and Bakt_806R [48]. The V3-V5 region is 464 base pairs (bp) long, with forward and reverse reads overlapping by 136 bp, using 2 X 300 bp paired-end sequencing [110] (http://probebase.csb.univie.ac.at). Primer sequences include overhang adapter sequences for library preparation (forward primer 5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG - 3' and reverse primer 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC - 3'). For quality control, the PCR product was verified using agarose gel electrophoresis to check for appropriate size bands, and concentration measurements were made after the initial 16S rRNA PCR, the indexing PCR, and normalization steps. DNA concentration was measured using SpextraMax Accuclear Nano dsDNA Assay Bulk Kit (Part# R8357#, Lot 215737, Molecular Devices LLC. Sunnyvale CA, USA) and fluorescent measurements were made with a Molecular Devices SpectraMax M2 spectraflourometer (Molecular Devices LLC. Sunnyvale CA, USA).

Initial PCR products were purified using AMPure XP beads (Beckman Coulter Genomics, Danvers, MA) following the manufacturer's protocol. After purification, the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). Prior to pooling purified sample concentration was normalized using SequalPrep Normalization Plate Kit (Catalog n. A10510-01, Invitrogen Corp., Carlsbad, CA), according to the manufacturer's protocol. Pooled library concentration was checked using the Qubit dsDNA HS Assay Kit (Part# Q32851, Lot# 1735902, ThermoFisher, Waltham, MA USA). Due to the low pooled amplicon library DNA concentration, a modified protocol for low concentration libraries was used. The library was run on an Illumina MiSeq, and base calls were made using Illumina Real Time Analysis Software version 1.18.54. Sequencing data quality control metrics for the 384 fastq sequence files (192 samples with forward and reverse reads) were computed using the Bioconductor `Rqc` package [93, 39].

3.3.4 Sequence Processing

Sequence data were processed using four bioinformatic pipelines: a *de-novo* clustering method - Mothur [89], an open-reference clustering method - QIIME [21], and a sequence inference methods - DADA2 [19], and unclustered sequences as a control. The code used to run the bioinformatic pipelines is available at `https://github.com/nate-d-olson/mgtst_pipelines`.

The Mothur pipeline follows the developers MiSeq SOP [89, 51]. The pipeline was run using Mothur version 1.37 (`http://www.mothur.org/`) As we sequenced a larger 16S rRNA region, with smaller overlap between the forward and reverse reads, than the 16S rRNA region the SOP was designed. Pipeline parameters were modified to account for the difference in overlap are noted for individual steps below. The Makefile and scripts used to run the mothur pipeline are available `https://github.com/nate-d-olson/mgtst_pipelines/blob/master/code/mothur`. The Mothur pipeline included an initial preprocessing step where the forward and reverse reads are trimmed and filtered using base quality scores merged into contigs. The following parameters were used for the initial contig filtering, no ambiguous bases, max contig length of 500 bp, and max homopolymer length of 8 bases. For the initial read filtering and merging step, low-quality reads were identified and filtered from the dataset based on the presence of ambiguous bases, failure to align to the SILVA reference database (V119, `https://www.arb-silva.de/`) [79], and identification as chimeras. Prior to alignment, the SILVA reference multiple sequence alignment was trimmed to the V3-V5 region, positions 6,388 and 25,316. Chimera filtering was performed using UChime (version v4.2.40) without a reference database [32]. OTU clustering was performed using the OptiClust algorithm with a clustering threshold of 0.97 [103]. The RDP classifier implemented in mothur was used for taxonomic classification against the mothur provided version of the RDP v9 training set [101].

The QIIME open-reference clustering pipeline for paired-end Illumina data was performed according to the online tutorial (`http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb`) using QIIME version 1.9.1 [21]. Briefly, the QIIME pipeline uses fastq-join (version 1.3.1) to merge paired-end reads [3] and the Usearch algorithm [29] with Greengenes database version 13.8 with a 97% similarity threshold [28] was used for open-reference clustering.

DADA2, an R native pipeline was also used to process the sequencing data [19]. The pipeline includes a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP naïve Bayesian classifier [101] and the SILVA database V123 provided by the DADA2 developers [79, `https://benjjneb.github.io/dada2/training.html`].

The unclustered pipeline was based on the mothur *de-novo* clustering pipeline, where the paired-end reads were merged, filtered, and then dereplicated. Reads were aligned to the reference Silva alignment (V119, `https://www.arb-silva.de/`), and reads failing alignment were excluded from the dataset. Taxonomic classification of the unclustered sequences was performed using the same RDP classifier implemented in mothur used for the *de-novo* pipeline. To limit the size of the dataset the most abundant 40,000 OTUs (comparable to the mothur dataset), across all samples, were used as the unclustered dataset.

3.3.5 Titration Proportion Estimates

The following linear model (3.2) was used to infer the proportion of prokaryotic DNA in each titration, θ. Where $\mathbf{Q}_i$ is a vector of titration i feature relative abundance estimates and $\mathbf{Q}_{pre}$ and $\mathbf{Q}_{post}$ are vectors of feature relative abundance estimates for the unmixed PRE and POST samples. Average PCR replicate relative abundance values were calculated using a negative binomial model.

$$\mathbf{Q}_i = \theta_i(\mathbf{Q}_{post} - \mathbf{Q}_{pre}) + \mathbf{Q}_{pre} \tag{3.2}$$

To fit the model to prevent uninformative and low abundance features from biasing θ estimates only informative features meeting the following criteria were used Features included in the model were observed in at least 14 of the 28 total titration PCR replicates (4 replicates per 7 titrations), demonstrated greater than 2-fold difference in relative abundance between the PRE and POST samples, and were present in either all four or none of the PRE and POST PCR replicates.

16S rRNA sequencing count data is known to have a non-normal mean-variance relationship resulting in poor model fit for standard linear regression [62]. Generalized linear models provide an alternative to standard least-squares regression. The above model is additive and therefore unable to directly infer θ_i in log-space. To address this issue, we fit the model using a standard least-squares regression then obtained non-parametric 95 % confidence intervals for the θ estimates by bootstrapping with 1000 replicates.

3.3.6 Qualitative Assessment

Our qualitative measurement assessment evaluated features only observed in unmixed samples (PRE or POST), *unmixed-specific*, or titrations,*titration-specific*. *Unmixed-* or *titration-specific* features are due to differences in sampling depth (number of sequences) between the unmixed samples and titrations, artifacts of the feature inference process, or PCR/sequencing artifacts. Measurement process artifacts should be considered false positives or negatives. Hypothesis tests were used to determine if differences in sampling depth could account for *unmixed-specific* and *titration-specific* features. p-values were adjusted for multiple comparisons using the Benjamini & Hochberg method [8]. For *unmixed-specific* features, the binomial test was used to evaluate if true feature relative abundance is less than the expected relative abundance. A binomial test could not be used to evaluate *titration-specific* features, as the hypothesis would be formulated as such. Given observed counts and the titration total feature abundance, the true feature relative abundance is equal to 0. As non-zero counts were observed the true feature proportion is non-zero, and the test always fails. Therefore, we formulated a Bayesian hypothesis test for *titration-specific* features.

A Bayesian hypothesis test was used to evaluate if the true feature proportion is less than the minimum detected proportion. The Bayesian hypothesis test was formulated using equation (3.3). Which when assuming equal priors, $P(\pi < \pi_{min}) =$

$P(\pi \geq \pi_{min})$, reduces to (3.4). For equations (3.3) and (3.4) π is the true feature proportion, π_{min} is the minimum detected proportion, C is the expected feature counts, and C_{obs} is the observed feature counts. Simulation was used to generate possible values of C, assuming C has a binomial distribution given the observed sample total feature abundance, and a uniform probability distribution for π between 0 and 1. π_{min} was calculated using the mixture equation (3.1) where $q_{pre,j}$ and $q_{post,j}$ are $min(\mathbf{Q}_{pre})$ and $min(\mathbf{Q}_{post})$ across all features for a subject and pipeline. Our assumption is that π is less than π_{min} for features not observed in unmixed samples due to random sampling.

$$\begin{aligned} p &= P(\pi < \pi_{min} | C \geq C_{obs}) \\ &= \frac{P(C \geq C_{obs} | \pi < \pi_{min}) P(\pi < \pi_{min})}{P(C \geq C_{obs} | \pi < \pi_{exp}) P(\pi < \pi_{min}) + P(C \geq C_{obs} | \pi \geq \pi_{min}) P(\pi \geq \pi_{min})} \end{aligned} \tag{3.3}$$

$$p = \frac{P(C \geq C_{obs} | \pi < \pi_{min})}{P(C \geq C_{obs})} \tag{3.4}$$

3.3.7 Quantitative Assessment

Quantitative assessment compared observed relative abundance and log fold-changes to expected values derived from the titration experimental design. Feature average relative abundance across PCR replicates was calculated using a negative binomial model, and used as observed relative abundance values (obs) for the relative abundance assessment. Average relative abundance values were used to reduce PCR replicate outliers from biasing the assessment results. Equation (3.1) and inferred θ values were used to calculate the expected relative abundance values (exp). Relative abundance error rate is defined as $|exp - obs|/exp$.

We developed bias and variance metrics to assess feature performance. The feature-level bias and variance metrics were defined as the median error rate and robust coefficient of variation ($RCOV = IQR/median$) respectively. Mixed-effects models were used to compare feature-level error rate bias and variance metrics across pipelines with subject as a random effect. Extreme feature-level error rate bias and variance metric outliers were observed, these outliers were excluded from the mixed effects model to minimize biases due to poor model fit and were characterized independently.

Log fold-change between samples in the titration series including PRE and POST were compared to the expected log fold-change values to assess differential abundance log fold-change estimates. Log fold-change estimates were calculated using EdgeR [84, 58]. Expected log fold-change for feature j between titrations l and m is calculated using equation (3.5), where θ is the proportion of POST bacterial DNA in a titration, and q is feature relative abundance. For features only present in PRE samples the expected log fold-change is independent of the observed counts for the unmixed samples and is calculated using (3.6). Due to a limited number of *PRE-specific* features, both *PRE-specific* and *PRE-dominant* features were used in the differential abundance assessment. *PRE-specific* features were defined as features observed in all four PRE PCR replicates and not observed in any of the

Table 3.1: Summary statistics for the different bioinformatic pipelines. DADA2 is a denoising sequence inference pipeline, QIIME is an open-reference clustering pipeline, and mothur is a de-novo clustering pipeline. No template controls were excluded from summary statistics. Sparsity is the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2) in the count. Sample coverage is the median and range (minimum-maximum) per sample total abundance. Drop-out rate is the proportion of reads removed while processing the sequencing data for each bioinformatic pipeline.

Pipelines	Features	Sparsity	Total Abundance	Drop-out Rate
DADA2	3144	0.93	68649 (1661-112058)	0.24 (0.18-0.59)
Mothur	38469	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
QIIME	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

POST PCR replicates and *PRE-dominant* features were also observed in all four PRE PCR replicates and observed in one or more of the POST PCR replicates with a log fold-change between PRE and POST samples greater than 5.

$$logFC_{lm,j} = \log_2\left(\frac{\theta_l q_{post,j} + (1-\theta_l)q_{pre,i}}{\theta_m q_{post,j} + (1-\theta_m)q_{pre,j}}\right) \tag{3.5}$$

$$logFC_{lm,i} = log_2\left(\frac{1-\theta_l}{1-\theta_m}\right) \tag{3.6}$$

3.4 Results

3.4.1 Dataset characteristics

We first characterize the number of reads per sample and base quality score distribution. The number of reads per sample and distribution of base quality scores by position was consistent across subjects (Fig. 3.2). Two barcoded experimental

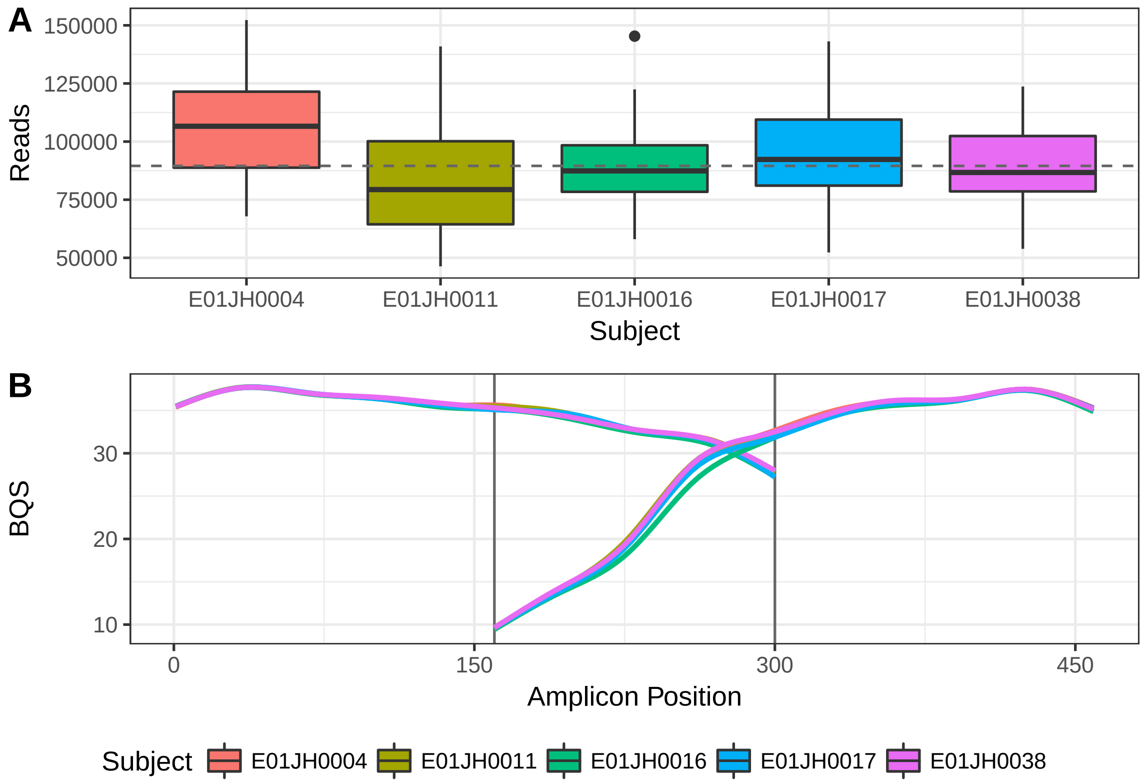


Figure 3.2: Sequence dataset characteristics. (A) Distribution in the number of reads per barcoded sample (Library Size) by individual. The dashed horizontal line indicates overall median library size. Excluding one PCR replicate from subject E01JH0016 titration 5 that had only 3,195 reads. (B) Smoothing spline of the base quality score (BQS) across the amplicon by subject. Vertical lines indicate approximate overlap region between forward and reverse reads. Forward reads go from position 0 to 300 and reverse reads from 464 to 164.

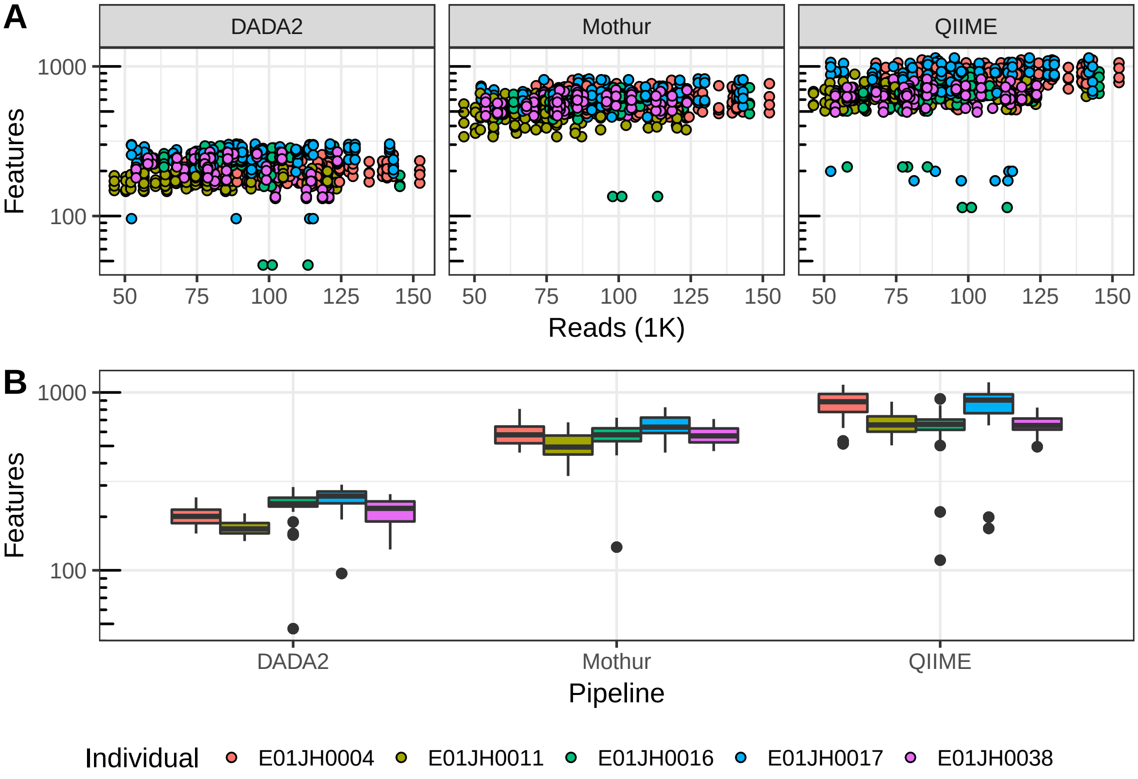


Figure 3.3: Relationship between the number of reads and features per sample by bioinformatic pipeline. (A) Scatter plot of observed features versus the number of reads per sample. (B) Observed feature distribution by pipeline and individual. Excluding one PCR replicate from subject E01JH0016 titration 5 with only 3,195 reads, and the Mothur E01JH0017 titration 4 (all four PCR replicates), with 1,777 observed features.

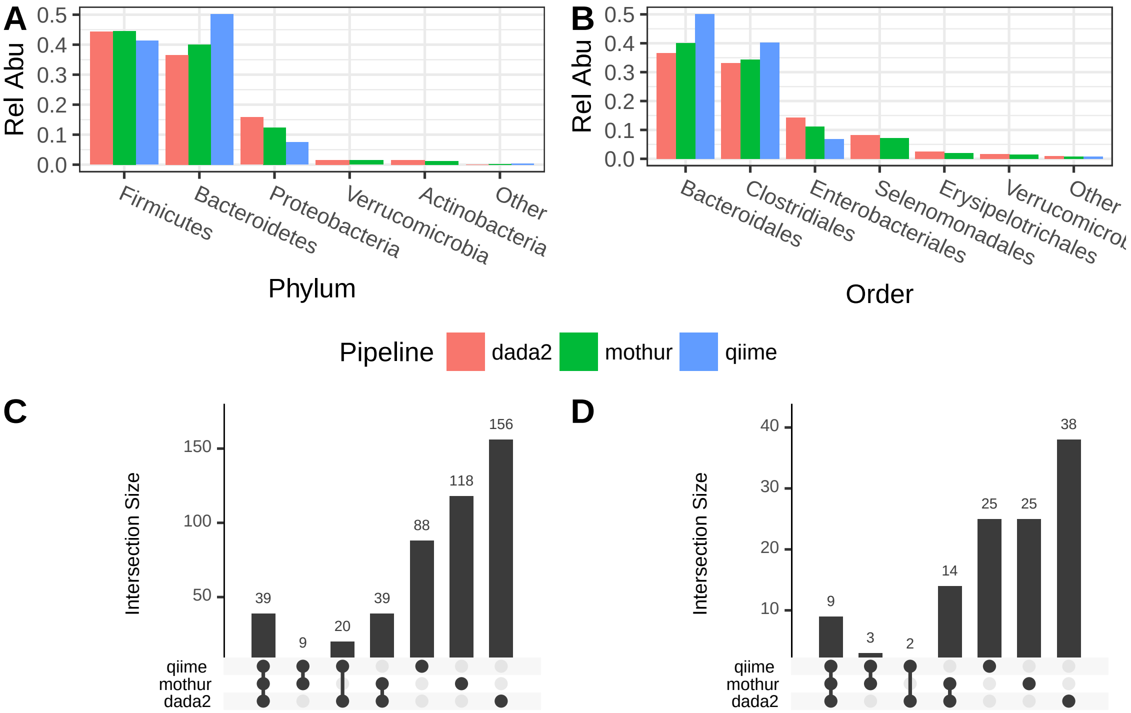


Figure 3.4: Comparison of dataset taxonomic composition across pipelines. Phylum (A) and Order (B) relative abundance by pipeline. Taxonomic groups with less than 1% total relative abundance were grouped together and indicated as other. Pipeline genus-level taxonomic assignment set overlap for the all features (C) and the upper quartile genera by relative abundance for each pipeline (D).

samples had less than 35,000 reads. The rest of the samples with less than 35,000 reads were no template PCR controls (NTC). Excluding the one failed reaction with 2,700 reads and NTCs, there were 8.9548×10^4 (3195-152267, median and range) sequnces per sample. The forward read has consistently higher base quality scores relative to the reverse read with a narrow overlap region with high base quality scores for both forward and reverse reads (Fig. 3.2B).

The resulting count tables generated using the four bioinformatic pipelines were characterized for number of features, sparsity, and filter rate(Table 3.1, Figs. 3.3B). The pipelines evaluated employ different approaches for handling low quality reads resulting in the large differences in drop-out rate and the fraction of raw sequences not included in the count table (Table 3.1). QIIME pipeline has the highest drop-out rate and number of features per sample but fewer total features than Mothur. The targeted amplicon region has a relatively small overlap region, 136 bp for 300 bp paired-end reads, compared to other commonly used amplicons [51, 100]. The high drop-off rate is due to low basecall accuracy at the ends of the reads especially the reverse reads resulting in a high proportion of unsuccessfully merged reads pairs (Fig. 3.2B). Furthermore increasing the drop-out rate, QIIME excludes singletons, OTUs only observed once in the dataset, to remove potential sequencing artifacts from the dataset. QIIME and DADA2 pipelines were similarly sparse (the fraction of zero values in count tables) despite differences in the number of features and drop-out rate. The expectation is that this mixture dataset will be less sparse relative to other datasets due to the redundant nature of the samples where 35 of the samples are derived directly from the other 10 samples, and four PCR replicates for each sample. With sparsity greater than 0.9 for the three pipelines it is unlikely that any of the pipelines successfully filtered out a majority of the sequencing artifacts.

The dataset taxonomic assignments also varied by pipeline (Fig. 3.4). Phylum and order relative abundance is similar across pipelines (Fig. 3.4A & B). Differences are attributed to different taxonomic classification methods and databases. The DADA2 and QIIME pipelines differed from Mothur and QIIME for Proteobacteria and Bacteriodetes. Regardless of threshold, for genus sets most genera were unique to individual pipelines (Fig. 3.4C & D). Sets with QIIME had the fewest genera, excluding the DADA2-QIIME set. QIIME pipeline was the only one to use the open-reference clustering and the Greengenes database. Mothur and DADA2 both used the SILVA dataset. The Mothur and DADA2 pipeline use different implmentations of the RDP naïve Bayesian classifier, which may be partially responsible for the mothur, unclustered, and DADA2 differences.

3.4.2 Titration Series Validation

To validate the two-sample titration dataset for use in abundance assessment we evaluated two assumptions about the titrations: 1. The samples were mixed volumetrically in a log_2 dilution series according to the mixture design. 2. The unmixed PRE and POST samples have the same proportion of prokaryotic DNA. To validate the sample volumetric mixing exogenous DNA was spiked into the unmixed samples before mixing and quantified using qPCR . To evaluate if the PRE and

POST samples had the same proportion of prokaryotic DNA total prokaryotic DNA in the titrations samples was quantified using a qPCR assay targeting the 16S rRNA gene.

3.4.2.1 Spike-in qPCR results

Titration series volumetric mixing was validated by quantify ERCC plasmids spiked into the POST samples using qPCR. The qPCR assay standard curves had a high level of precision with R^2 values close to 1 and amplification efficiencies between 0.84 and 0.9 for all standard curves indicating the assays were suitable for validating the titration series volumetric mixing (Table 3.2). For our log_2 two-sample-titration mixture design the expected slope of the regression line between titration factor and Ct is 1, corresponding to a doubling in template DNA every PCR cycle. The qPCR assays targeting the ERCCs spiked into the POST samples had R^2 values and slope estimates close to 1 (Table 3.2). Slope estimates less than one were attributed to assay standard curve efficiency less than 1 (Table 3.2). ERCCs spiked into PRE samples were not used to validate volumetric mixing as PRE sample proportion differences were too small for qPCR quantification. The expected C_t difference for the entire range of PRE concentrations in only 1. When considering the quantitative limitations of the qPCR assay these results confirm that the unmixed samples were volumetrically mixed according to the design.

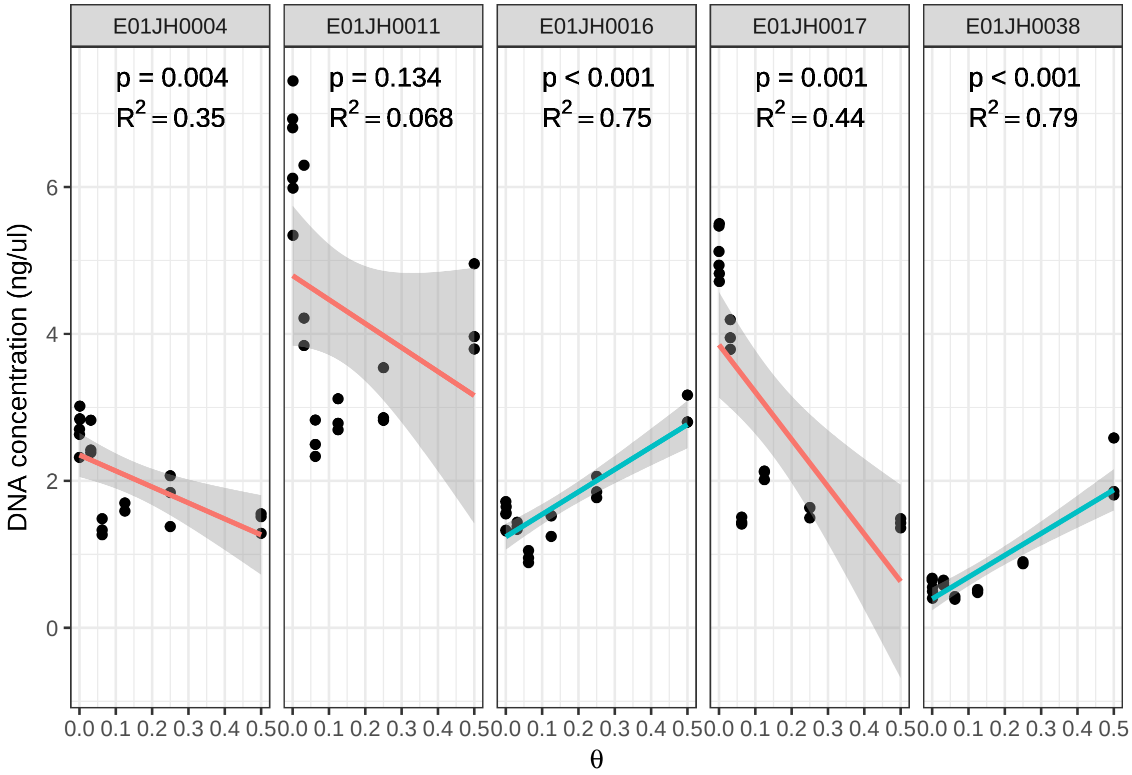


Figure 3.5: Prokaryotic DNA concentration (ng/ul) across titrations measured using a 16S rRNA qPCR assay. Separate linear models, Prokaryotic DNA concentration versus θ were fit for each individual, and R^2 and p-values were reported. Red lines indicate negative slope estimates and blue lines positive slope estimates. p-value indicates significant difference from the expected slope of 0. Multiple test correction was performed using the Benjamini-Hochberg method. One of the E01JH0004 PCR replicates for titration 3 ($\theta = 0.125$) was identified as an outlier, with a concentration of 0.003, and was excluded from the linear model. The linear model slope was still significantly different from 0 when the outlier was included.

Table 3.2: ERCC Spike-in qPCR assay information and summary statistics. ERCC is the ERCC identifier for the ERCC spike-in, Assay is TaqMan assay, and Length and GC are the size and GC content of the qPCR amplicon. The Std. R^2 and Efficiency (E) statistics were computed for the standard curves. R^2 and slope for titration qPCR results for the titration series.

Subject	ERCC	Assay	Length	Std. R^2	E	R^2	Slope
E01JH0004	012	Ac03459877-a1	77	0.9996	86.19	0.98	0.92
E01JH0011	157	Ac03459958-a1	71	0.9995	87.46	0.95	0.90
E01JH0016	108	Ac03460028-a1	74	0.9991	87.33	0.95	0.84
E01JH0017	002	Ac03459872-a1	69	0.9968	85.80	0.89	0.93
E01JH0038	035	Ac03459892-a1	65	0.9984	86.69	0.95	0.94

3.4.2.2 Bacterial DNA Concentration

The observed changes in prokaryotic DNA concentration across titrations indicate the proportion of bacterial DNA from the unmixed PRE and POST samples in a titration is inconsistent with the mixture design (Fig. 3.5). A qPCR assay targeting the 16S rRNA gene was used to quantify the concentration of prokaryotic DNA in the titrations. An in-house standard curve with concentrations of 20 ng/ul, 2ng/ul, and 0.2 ng/ul was used, with efficiency 91.49, and R^2 0.999. If the proportion of prokaryotic DNA is the same between PRE and POST samples the slope of the concentration estimates across the two-sample titration would be 0. For subjects where the proportion of prokaryotic DNA is higher in the PRE samples, the slope will be negative and positive when the proportion is higher for POST samples. The slope estimates are significantly different from 0 for all subjects excluding E01JH0011 (Fig. 3.5). These results indicate that the proportion of prokaryotic DNA is lower in POST when compared to the PRE samples for E01JH0004 and E01JH0017 and higher for E01JH0016 and E01JH0038.

3.4.2.3 Theta Estimates

To account for differences in the proportion of prokaryotic DNA in PRE and POST samples (Fig. 3.5) we inferred the proportion of POST sample prokaryotic DNA in a titration, θ, using the 16S rRNA sequencing data (Fig. 3.6). Overall the relationship between the inferred and mixture design θ values were consistent across pipelines but not subject whereas the 95% CI varied by both subject and pipeline. For study subjects E01JH0004, E01JH0011, and E01JH0016 the inferred and mixture design θ values were in agreement, in contrast, to study subjects E01JH0017 and E01JH0038. For E01JH0017 the inferred values were consistently less than the mixture design values. Whereas for E01JH0038 the inferred values were consistently greater than the mixture design values. These results were consistent with the qPCR prokaryotic DNA concentration results with significantly positive slopes for E01JH0004 and E01JH0016 and a significantly negative slope for E01JH0038 (Fig. 3.5).

3.4.3 Measurement Assessment

Next, we assessed the qualitative and quantitative nature of 16S rRNA measurement process using our two-sample titration dataset. For the qualitative assessment, we analyzed the relative abundance of features only observed in the unmixed samples or titrations which are not expected given the titration experimental design. The quantitative assessment evaluated relative and differential abundance estimates.

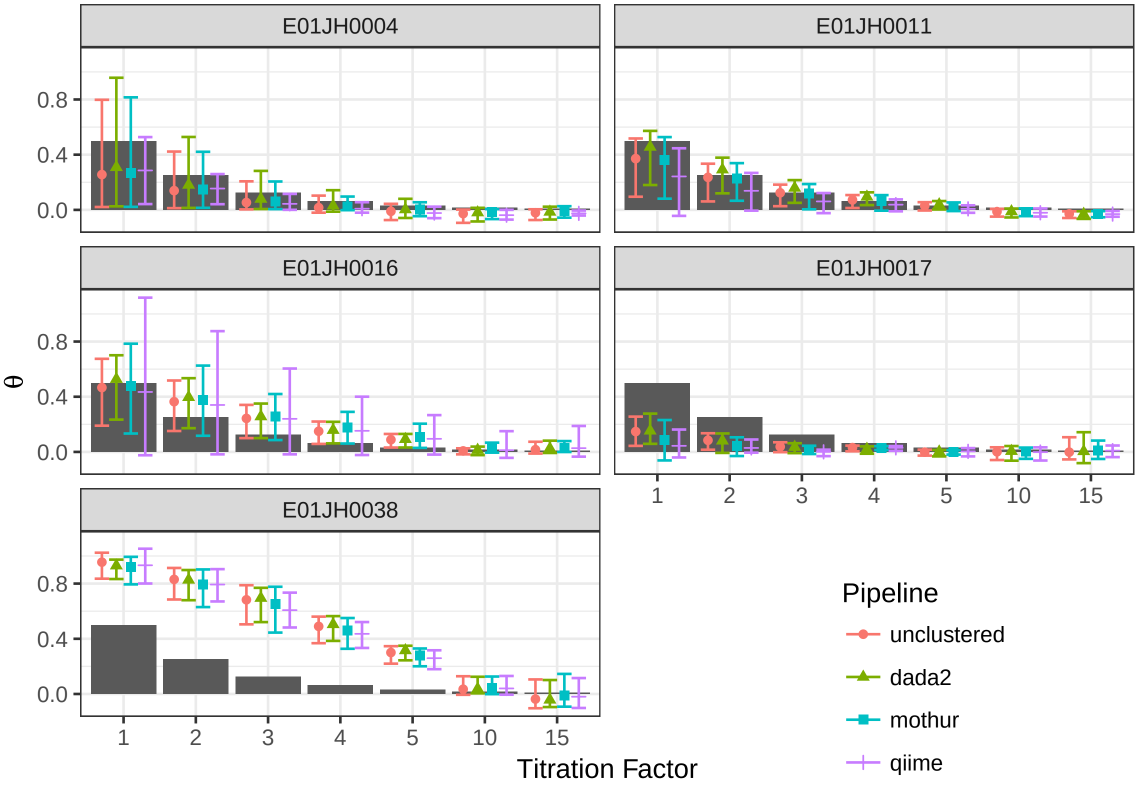


Figure 3.6: Theta estimates by titration, biological replicate, and bioinformatic pipeline. The points indicates mean estimate of 1000 bootstrap theta estimates and errorbars 95% confidence interval. The black bar indicate expected theta values. Theta estimates below the expected theta indicate that the titrations contain less than expected bacterial DNA from the POST sample. Theta estimates greater than the expected theta indicate the titration contains more bacterial DNA from the PRE sample than expected.

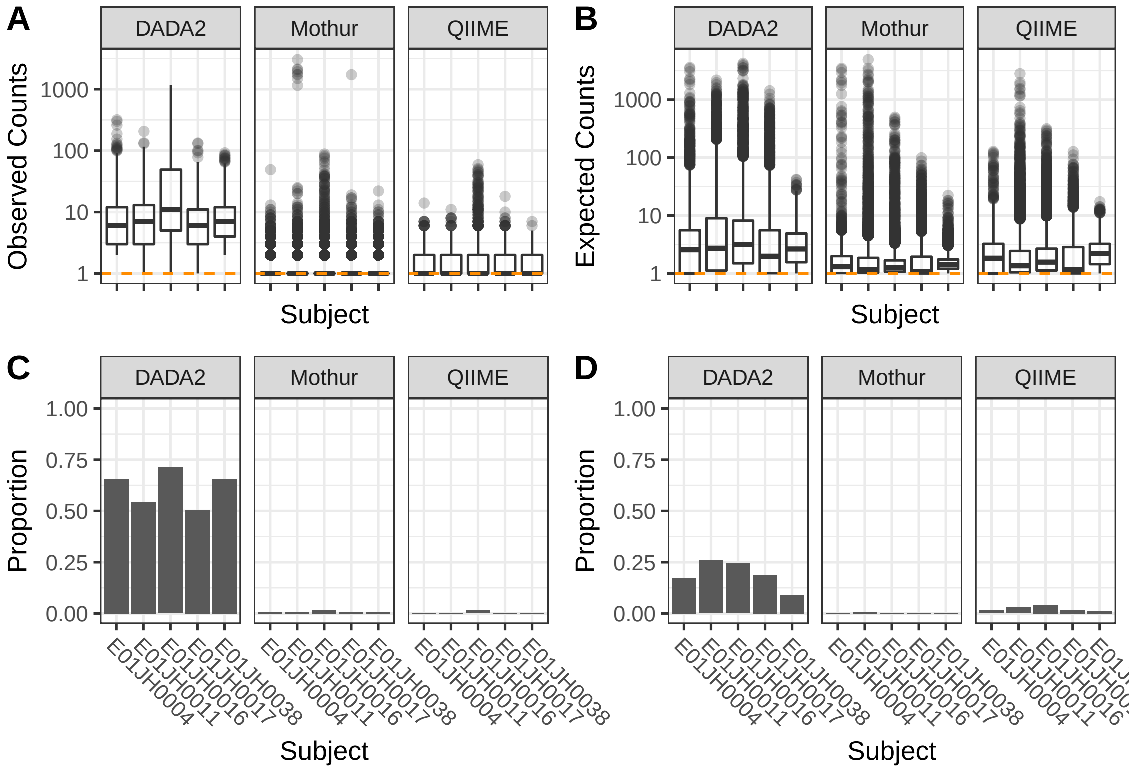


Figure 3.7: Distribution of (A) observed count values for titration-specific features and (B) expected count values for unmixed-specific features by pipeline and individual. The orange horizontal dashed line indicates a count value of 1. (C) Proportion of unmix-specific features and (D) titration-specific features with an adjusted p-value < 0.05 for the Bayesian hypothesis test and binomial test respectively. We failed to accept the null hypothesis when the p-value < 0.05, indicating that the discrepancy between the feature only being observed in the titrations or unmixed samples cannot be explained by sampling alone.

3.4.3.1 Qualitative Assessment

Unmixed- and titration-specific features were observed for all pipelines (titration-specific: Fig. 3.7A, unmixed-specific: Fig. 3.7B). For mixture datasets the low abundance features present only in the unmixed samples and mixtures are expected due to random sampling. For our two-sample titration dataset there were unmixed-specific features with expected counts that could not be explained by sampling alone for all individuals and bioinformatic pipelines (Fig. 3.7C). However, the proportion

Table 3.3: Maximum feature-level error rate bias (median error rate) and variance (robust COV) by pipeline and individual.

Metric	Pipeline	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
Bias	dada2	2.37	2.55	17.03	4.34	0.66
	mothur	5.30	6.76	19.24	4.15	1.93
	qiime	3.99	6.43	8.83	4.80	1.09
	unclustered	6.45	7.24	16.85	4.37	1.91
Variance	dada2	4.60	8.96	7.36	5.91	6.71
	mothur	4.71	7.35	3.71	5.70	8.01
	qiime	4.40	22.57	4.46	17.10	7.91
	unclustered	7.06	10.30	16.94	8.07	6.00

of unmixed-specific features that could not be explained by sampling alone varied by bioinformatic pipeline. DADA2 had the highest rate of unmixed-specific features not explained by sampling whereas QIIME had the lowest rate. Consistent with the distribution of observed counts for titration-specific features more of the DADA2 features could not be explained by sampling alone compared to the other pipelines (Fig. 3.7D). Overall, DADA2 resulted in the largest number of observed features inconsistent with the titration experiment design, while the same phenomenon is significantly reduced in the other pipelines.

3.4.3.2 Quantitative Assessment

For the relative abundance assessment, I evaluated the consistency of the observed and expected relative abundance estimates for a feature and titration as well as feature-level bias and variance. The PRE and POST estimated relative abundance and inferred θ values were used to calculate titration and feature level error rates. Unclustered pipeline θ estimates were used to calculate the error rates for all

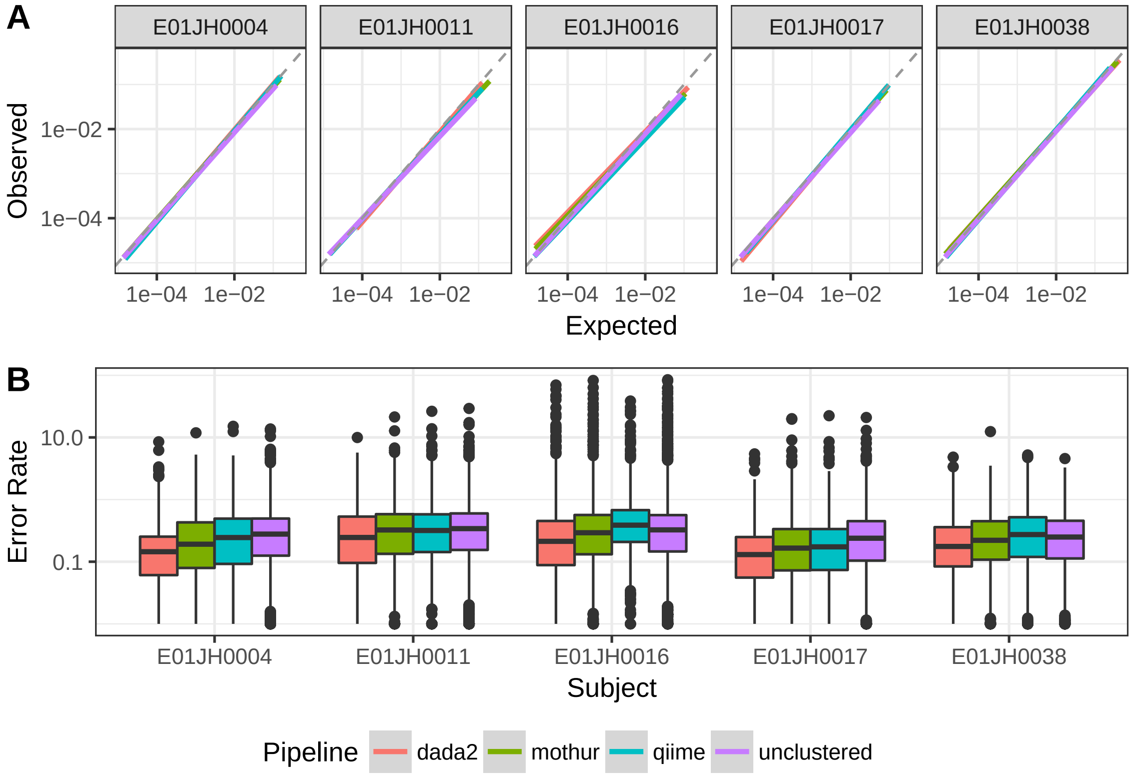


Figure 3.8: Relative abundance assessment. (A) A linear model of the relationship between the expected and observed relative abundance. The dashed grey line indicates expected 1-to-1 relationship. The plot is split by individual and color is used to indicate the different bioinformatic pipelines. A negative binomial model was used to calculate an average relative abundance estimate across the four PCR replicates. Points with observed and expected relative abundance values less than 1/median library size were excluded from the data used to fit the linear model. (B) Relative abundance error rate distribution by individual and pipeline.

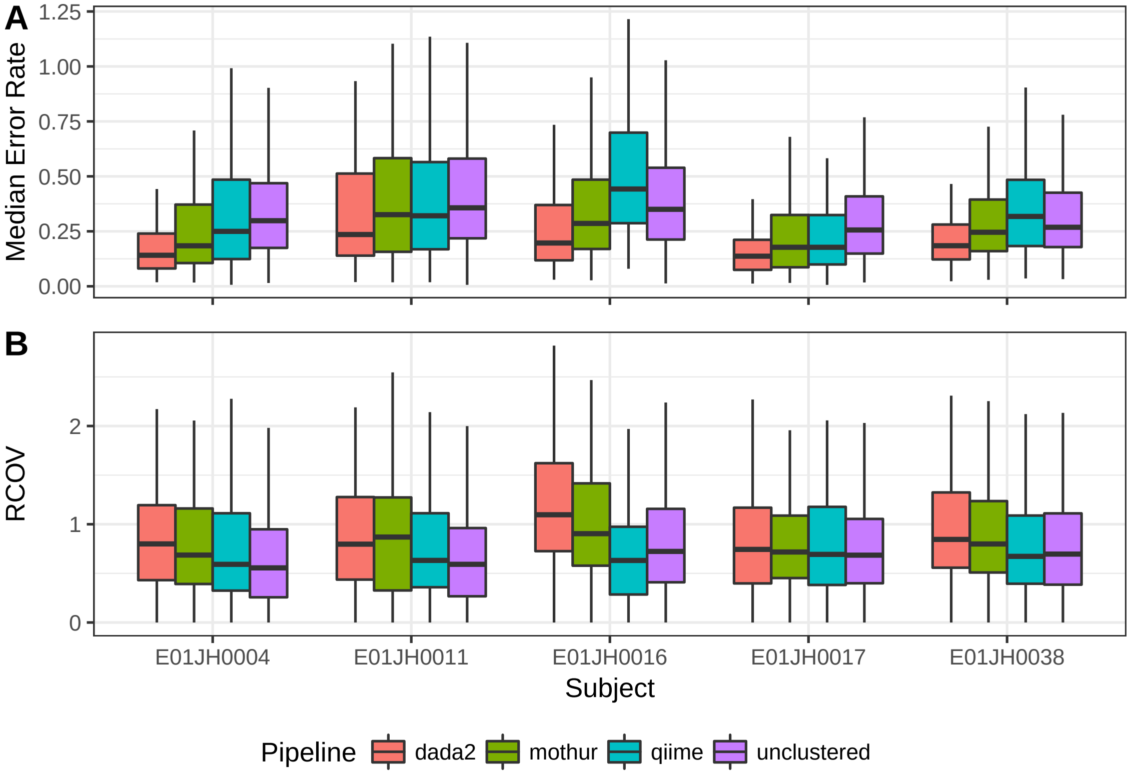


Figure 3.9: Comparison of pipeline relative abundance assessment feature-level error metrics. Distribution of feature-level relative abundance (A) bias metric - median error rate and (B) variance - robust coefficient of variation ($RCOV = (IQR)/|median|$) by individual and pipeline. Boxplot outliers, $1.5 \times IQR$ from the median were excluded from the figure to prevent extreme metric values from obscuring metric value visual comparisons.

pipelines to prevent over-fitting. Only features observed in all PRE and POST PCR replicates and PRE and POST specific features were included in the analysis (Table 3.3). PRE and POST specific features were defined as present in all four PCR replicates of the PRE or POST PCR replicates, respectively, but none of the PCR replicates for the other unmixed samples. There is lower confidence in PRE or POST feature relative abundance when the feature is not observed in some of the 4 PCR replicates, therefore these features were not included in the error analysis. Overall, agreement between the inferred and observed relative abundance was high for all individuals and bioinformatic pipelines (Fig. 3.8A). The error rate distribution was similarly consistent across pipelines, including long tails (Fig. 3.8B)

To assess quantitative accuracy I compared the feature-level relative abundance error rate bias (median error rate, Fig. 3.9A) and variance ($RCOV = (IQR)/|median|$ Fig. 3.9B) across pipelines and individuals using mixed effects models. Large bias and variance values were observed for all pipelines (Table 3.3). Features with large bias and variance metrics (outliers), defined as $1.5 \times IQR$ from the median. To prevent the outliers from biasing the comparison they were not included in the dataset used to fit the mixed effects model. Multiple comparisons test (Tukey) was used to test for significant differences in feature-level bias and variance between pipelines. A one-sided alternative hypothesis was used to determine which pipelines had a smaller, feature-level error rate. The Mothur, DADA2, and QIIME feature-level bias were all significantly different from each other ($p < 1 \times 10^{-8}$). DADA2 had the lowest mean feature-level bias (0.2), followed by Mothur (0.28), with QIIME having the highest bias (0.33) (3.9B). Large variance metric values

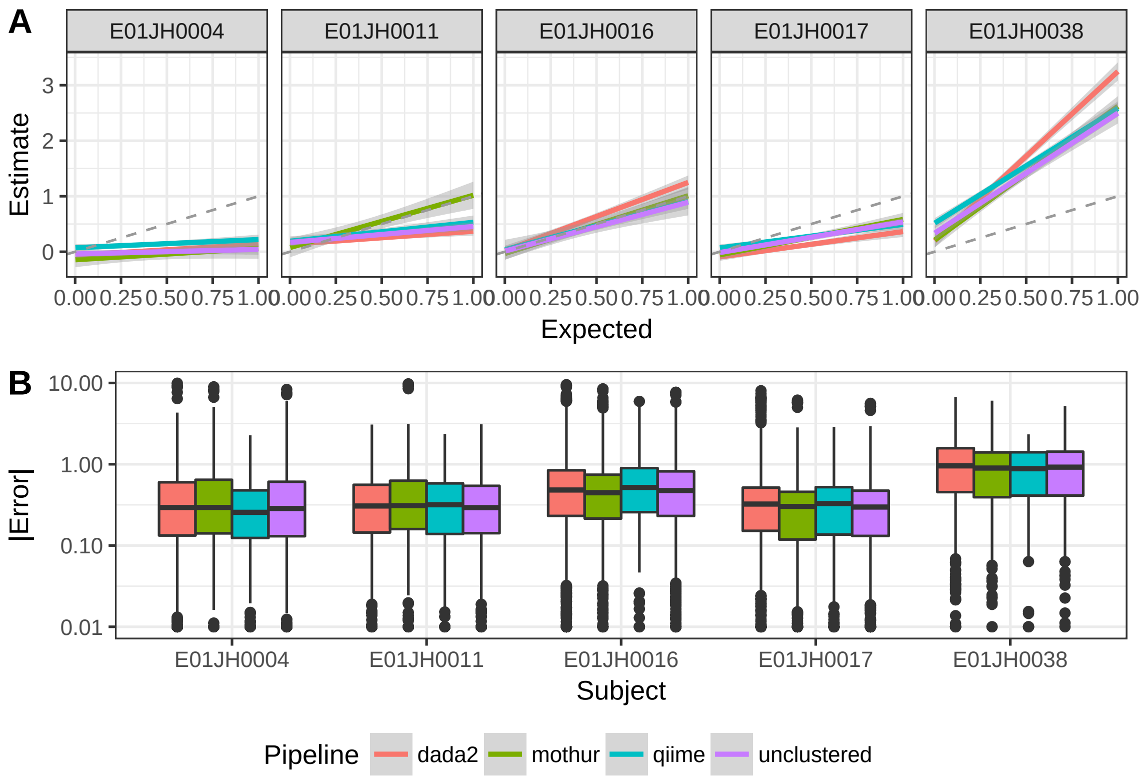


Figure 3.10: (A) Linear model or the relationship between log fold-change estimates and expected values for PRE-specific and PRE-dominant features by pipeline and individual, line color indicates pipelines. Dashed grey line indicates expected 1-to-1 relationship between the estimated and expected log fold-change. (B) Log fold-change error (|exp-est|) distribution by pipeline and individual.

were observed for all individuals and pipelines (Table 3.3). The feature-level variance was not significantly different between pipelines, Mothur = 0.83, QIIME = 0.71 and DADA2 = 1 (Fig. 3.9B). I evaluated whether poor feature-level relative abundance metrics can be attributed to specific taxonomic groups or phylogenetic clades. While a significant overall phylogenetic signal was detected for both the bias and variance metric, I was unable to identify specific taxonomic groups or phylogenetic clades exceedingly poor performance in our assessment.

The agreement between the log-fold change estimates and expected values were individual specific and consistent across pipelines (Fig. 3.10A). The individual spe-

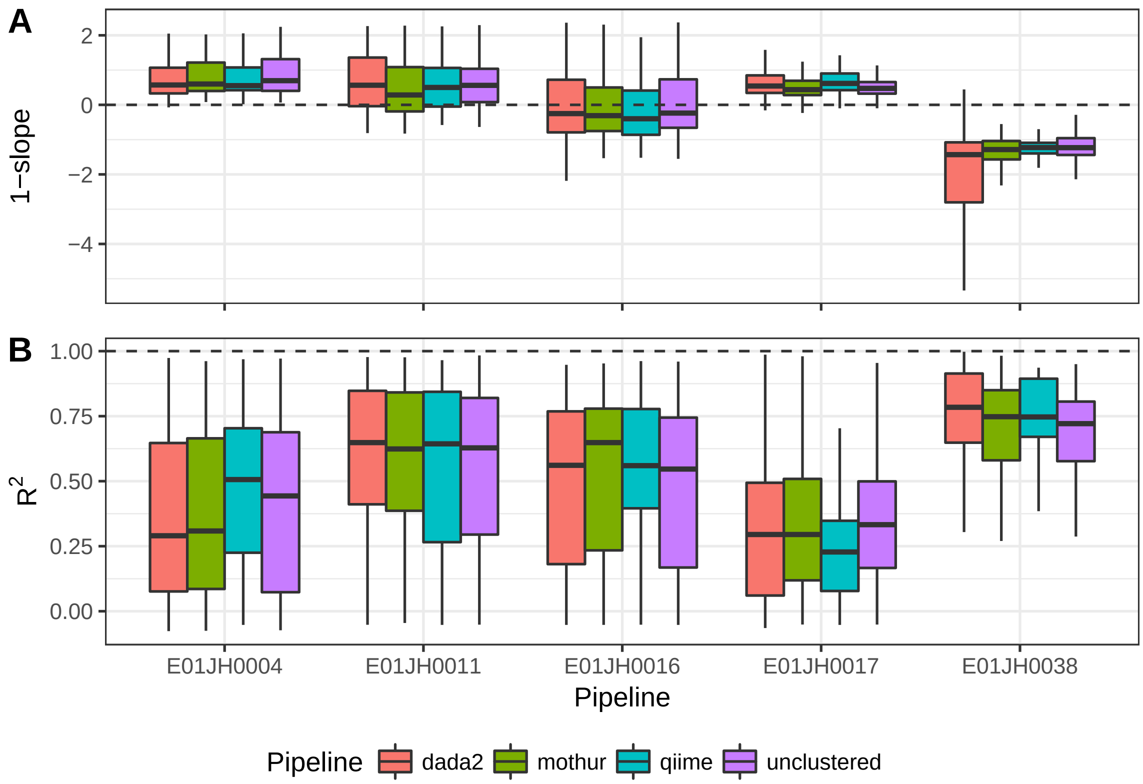


Figure 3.11: Feature-level log-fold change error bias (A) and variance (B) metric distribution by subject and pipeline. The bias ($1 - slope$) and variance (R^2) metrics are derived from the linear model fit to the estimated and expected log fold-change values for individual features. Boxplot outliers, 1.5 × IQR from the median were excluded from the figure to prevent extreme metric values from obscuring metric value visual comparisons.

cific effect was attributed to the fact that unlike the relative abundance assessment the inferred θ values were not used to calculate the expected values. The inferred θ values were not used to calculate the expected values as I wanted to include all of the titrations and the θ estimates for the higher titrations were not monotonically decreasing and therefore resulted in unrealistic expected log fold-change values, e.g., negative log-fold changes for PRE specific features. The log-fold change estimates and expected values were consistent across pipelines with one notable exception. For E01JH0011 the Mothur log fold-change estimates were in better agreement with the expected value compared to the other pipelines. However, as θ was not corrected for differences in the proportion of prokaryotic DNA between the unmixed PRE and POST samples I am unable to say whether Mothur's performance was better than the other pipelines.

The log fold-change error distribution was consistent across pipelines (Fig. 3.10B). There was a long tail of high error features in the error distribution for all pipelines and individuals. The log fold-change estimates responsible for the long tail could not be attributed to specific titration comparisons. Additionally, I compared the log-fold change error distribution for log-fold change estimates using different normalization methods. The error rate distributions, including the long tails, were consistent across normalization methods. Furthermore, as the long tail was observed for the unclustered data as well, the log-fold change estimates contributing to the long tail are likely due to a bias associated with the molecular laboratory portion of the measurement process and not the bioinformatic pipelines. Based on exploratory analysis of the relationship between the log fold-change estimates and expected values for individual features indicated that the long tails were attributed to feature specific performance.

Feature-level log fold-change bias and variance metrics were used to compare pipeline performance (Fig. 3.10). Feature-level bias and variance metrics are defined as the $1 - slope$ and R^2 for linear models of the estimated and expected log fold-change for individual features and all titration comparisons. For the bias metric, $1 - slope$, the desired value is 0 (i.e., log fold-change estimate = log fold-change expected), with negative values indicating the log-fold change was consistently underestimated and positive values consistently overestimated. The linear model R^2 value was used to characterize the feature-level log fold-change variance as it indicates how consistent the relationship between log fold-change estimates and expected values is across titration comparisons. To compare bias and variance metrics across pipelines mixed-effects models were used. The log fold-change bias and variance metrics were not significantly different between pipelines (Bias: F = 0, 2.51, p = 0.99, 0.08, 3.10B, Variance: F = 47.39, 0.23, p = 0, 0.8, Fig. 3.10C). Next, I evaluated whether poor feature-level metrics could be attributed to specific clades for taxonomic groups. Similar to the relative abundance estimate, while a phylogenetic signal was detected for both the bias and variance metrics, I was unable to identify specific taxonomic groups or phylogenetic clades that performed poorly in our assessment.

3.5 Discussion

We assessed the quantitative and qualitative characteristics of count tables generated using different bioinformatic pipelines and 16S rRNA marker-gene survey mixture dataset. The mixture dataset followed a two-sample titration mixture design, where DNA collected before and after exposure to pathogenic *Escherichia coli* from five vaccine trial participants (subjects) were mixed following a log_2 dilution series (Fig. 3.1). Qualitative count table characteristics were assessed using relative abundance information for features observed only in titrations and unmixed samples. We quantitatively assed count tables by comparing feature relative and differential abundance to expected values.

3.5.1 Count Table Assessment Demonstration

We demonstrated our novel assessment approach by evaluating count tables generated using different bioinformatic pipelines, QIIME, Mothur, and DADA2. The Mothur pipeline uses *de novo* clustering for feature inference [103, 89]. Pairwise distances used in clustering are calculated using a multiple sequence alignment. The quality filtered paired-end reads are merged into contigs. The pipeline the aligns contigs to a reference multiple sequence alignment and removes uninformative positions in the multiple sequence alignment. The QIIME pipeline uses open-reference clustering where merged paired-end reads are first assigned to reference cluster centers [82, 21]. Next QIIME clusters unassigned reads *de novo*. Unlike Mothur, the QIIME clustering method uses pairwise sequence distances calculated from pairwise sequence alignments. As a result, the QIIME pairwise distances are calculated using the full ~436 bp sequences whereas Mothur pairwise distances were calculated using a 270 bp multiple sequence alignment. The DADA2 pipeline uses a probability model and maximization expectation algorithm for feature inference [19]. Unlike distance-based clustering methods employed by the Mothur and QIIME pipelines, DADA2 parameters determine if low abundance sequences are grouped with a higher abundance sequence. As a control, we compared our quantitative assessment results for the three pipelines to a count table of unclustered features. The unclustered features were generated using the Mothur pipeline preprocessing methods.

3.5.1.1 Quantitative Assesssment

While the relative abundance bias metric was significantly different between pipelines overall, pipeline choice had minimal impact on the quantitative assessment results when accounting for subject-specific effects. Outlier features, those with extreme quantitative analysis bias and variance metrics, were observed for all pipelines and both relative and differential abundance assessments. Outlier features could not be attributed to bioinformatic pipelines and are likely due to biases in the molecular biology part of the measurement process. Outlier features are not likely a pipeline artifact as they were observed in count tables generated using the unclustered pipeline as well as standard bioinformatic pipelines. We were unable to attribute outlier features to relative abundance values, log fold-change between unmixed samples, and sequence GC content. Features with extreme metric values were not limited to any specific taxonomic group or phylogenetic clade. PCR amplification is a well-known bias molecular biology component of the measurement process. Mismatches in the primer binding regions impact PCR efficiency and are a potential cause for poor feature-specific performance [109]. Additional research is needed before outlier features are attributed to mismatches in the primer binding regions.

3.5.1.2 Qualitative Assessment

The qualitative assessment evaluated whether features only observed in unmixed samples or titrations could be explained by sampling alone. Features present only in the titrations or unmixed samples not due to random sampling are bioinformatic pipeline artifacts. These artifacts can be categorized as false negative or false positive features. A false negative occurs when a lower abundance sequence representing an organism within the sample is clustered with a higher abundance sequence from a different organism. False positives are sequencing or PCR artifacts not appropriately filtered or assigned to an appropriate feature by the bioinformatic pipeline.

Count table sparsity, the proportion of zero-valued cells, provides additional insight into the qualitative assessment results. A high rate of false negative features is a potential explanation for the DADA2 count table's poor performance in the qualitative assessment and comparable sparsity to the other pipelines despite having significantly fewer features (Fig. 3.7, 3.1). The DADA2 feature inference algorithm may be aggressively grouping lower abundance true sequences with higher abundance sequences. As a result, the low abundance sequences are not present in samples leading to increased sparsity and higher abundance unmixed- and titration-specific features. Adjusting the DADA2 parameters, specifically the `OMEGA_A` parameter in `setDadaOpt`. Along these lines, the DADA2 documentation states that the default setting for `OMEGA_A` is conservative to prevent false positives at the cost of increasing false negatives [19].

False positive features provide an explanation for Mothur and QIIME pipelines having lower proportion of unmixed- and titration-specific features not explained by sampling but high sparsity (Fig. 3.7, 3.1). The statistical tests used to determine if the specific features could be explained by sampling only considers feature abundance. Therefore, the statistical test is not able to distinguish between true low abundance unmixed- and titration-specific features and low abundance sequence artifacts. Mothur and QIIME count tables have ten times and three times more features compared to DADA2, respectively (Table 3.1). While microbial abundance distributions are known to have long tails, it is likely that the observed sparsity is an artifact of the 16S rRNA sequencing measurement process. Similarly, significantly more features than expected are commonly observed for mock community benchmarking studies evaluating the QIIME and Mothur pipelines [51].

False positive features can be reduced, but not eliminated, using smaller amplicon and prevalence filtering. The 16S rRNA region sequenced in the study is larger than the region the *de-novo*, and open clustering pipelines were initially developed for, potentially explaining the higher than expected sparsity [51]. The larger region has a smaller overlap between the forward and reverse reads. As a result merging of the forward and reverse reads did not allow for the sequence error correction that occurs when a smaller amplicon is used. However, even when targeting smaller regions of the 16S rRNA gene both the *de-novo* (Mothur) and open-reference clustering (QIIME) pipelines produced count tables with significantly more features than expected in evaluation studies using mock communities. Prevalence filtering is used to exclude low abundance features, likely predominantly measurement artifacts [20]. For example, a study exploring the microbial ecology of the Red-necked stint *Calidris ruficollis*, a migratory shorebird, used a hard filter to validate their study conclusions are not biases by false positive features. The study authors compared results with and without prevalence filter ensuring that the study conclusions were not biased by using the arbitrary filter or including the low abundant features [83].

3.5.2 Using Mixtures to Assess 16S rRNA Sequencing

Mixtures of environmental samples have previously been used to assess RNAseq and microarray gene expression measurements. However, this is the first time mixtures have been used to assess microbiome measurement methods. Our mixture dataset allowed us to develop novel methods for assessing marker-gene-survey computational methods. Our quantitative assessment allowed for the characterization of relative abundance values using a dataset with a larger number of features and dynamic range compared to assessments using mock communities. As a result, we were able to identify previously unknown feature specific biases. Based on our study results additional experiments can be performed to identify the cause of these biases and develop appropriate methods to account for them. Based on our subject-specific results observation, we recommend that studies based on stool samples seeking inferences in a longitudinal series of multiple subjects carefully estimate bacterial DNA proportions and adjust inferences accordingly. Additionally, our qualitative assessment results, when combined with sparsity information provide a new method for evaluating how well bioinformatic pipelines account for sequencing artifacts without loss of true biological sequences.

There were also limitations using our mixture dataset. These limitations included: Lack of agreement between the proportion of unmixed samples titrations and the mixture design. The number of features used in the different analysis. These limitations are described below along with recommendations for addressing them in future studies.

Differences in the proportion of prokaryotic DNA in the samples used to generate the two-sample titrations series results in differences between the true mixture proportions and mixture design. We attempted to account for differences in mixture proportion from mixture design by estimating mixture proportions using sequence data. Similar to how the proportion of mRNA in RNA samples used in a previous mixture study. We were able to use an assay targeting the 16S rRNA gene to detect changes in the concentration of bacterial DNA across titration, but unable to quantify the proportion of bacterial DNA in the unmixed samples using qPCR data. Using the 16S sequencing data we inferred the proportion of bacterial DNA from the POST sample in each titration. However, the uncertainty and accuracy of the inference method are not known resulting in an unaccounted for error source.

A better method for quantifying sample bacterial DNA proportion or using samples with consistent proportions would increase the expected value and in-turn error metric accuracy. Limitations in the prokaryotic DNA qPCR concentration assay precision limit the suitability for use in mixture studies. Digital PCR provides a more precise alternative to qPCR and is, therefore, a more appropriate method. Alternatively using samples where the majority of the DNA is prokaryotic would minimize this issue. Mixtures of environmental samples can also be used to assess shotgun metagenomic methods as well. As shotgun metagenomics is not a targeted approach, differences in the proportion of bacterial DNA in a sample would not impact the assessment results in the same way as 16S rRNA marker-gene-surveys.

Using samples from a vaccine trial allowed for the use of a specific marker with an expected response, *E. coli*, during methods development. However, the high level of similarity between the unmixed samples resulted in a limited number of features that could be used in the quantitative assessment results. Using more diverse samples to generate mixtures would address this issue.

3.6 Conclusions

This two-sample-titration dataset can be used to evaluate and characterize bioinformatic pipelines and clustering methods. The sequence dataset presented in this study can be processed with any 16S bioinformatic pipeline. Our quantitative and qualitative assessment can then be performed on the count table and the results compared to those obtained using the pipelines included in this study. The threee pipelines we evaluated produced sets of features varying in total feature abundance, number of features per samples, and total features. The objective of any pipeline is to differentiate true biological sequences from artifacts of the measurement process. In general based on our evaluation results we recommend using for DADA2 for feature-level abundance analysis, e.g. differential abundance testing. While DADA2 performed poorly in our qualitative assessment, the pipeline had performed better in the quantitative assessment compared to the other pipelines. Additionally, the DADA2 poor qualitative assessment results due to false-negative features are unlikely to negatively impact feature-level abundance analysis, though additional research is needed to validate this claim. When determining which pipeline to use for a study, users should consider whether minimizing false positives (DADA2) or false negatives (Mothur) is more appropriate for their study objectives. When a sequencing dataset is processed using DADA2, the user can be more confident that an observed feature represents a member of the microbial community and not a measurement artifact. Pipeline parameter optimization could address DADA2 false-negative issue. For the Mothur and QIIME pipelines, prevalence filtering will reduce the number of false-positive features. Feature-level results for any 16S rRNA marker-gene survey should be interpreted with care, as the biases responsible for poor quantitative assessment are unknown. Addressing both of these issues requires advances in both the molecular biology and computational components of the measurement process.

CHAPTER 4

Diversity Assessment

Assessing the impact of sequencing characteristics on 16S rRNA marker-gene surveys beta-diversity analysis.

4.1 Abstract

Originally developed for macro-ecology, beta-diversity metrics are commonly used to assess overall community similarity between microbiome samples. The effects of sequencing depth and error rates on beta diversity calculations have not been thoroughly studied. In the following study, we evaluate the impact of sequence characteristics on beta-diversity analyses, and how well they are handled by different bioinformatic pipelines and normalization methods. We use a mixture dataset of stool samples from five vaccine trial participants, collected before and after exposure to a pathogen and mixed following a two-sample titration. The sequencing data were processed using six bioinformatics pipelines, including sequence inference, *de novo*, and reference based clustering approaches, along with nine normalization methods, including standard rarefaction approaches and numeric normalization techniques.

We assess (1) beta-diversity repeatability for PCR replicates across multiple sequencing libraries and runs, (2) the ability to differentiate groups of samples with varying levels of similarity and (3) differences in beta-diversity between biological and technical factors. The Mothur and DADA2 pipelines were more robust to sequencing errors compared to the other pipelines evaluated in the study. Out of the normalization methods compared in the study we suggest using total sum scaling for weighted metrics. Normalizing counts using rarefaction improved assessment results for unweighted metrics. Furthermore, we found normalization methods developed for microarray and RNA sequencing data, including trimmed mean of M values (TMM) and relative log expression (RLE), may not be appropriate for marker-gene survey beta-diversity analysis.

4.2 Introduction

Microbial communities are frequently characterized by targeting a marker-gene of interest (e.g., the 16S rRNA gene) for PCR amplification and high-throughput sequencing [34]. While these approaches have been successfully used to improve our understanding of microbiota taxonomy and diversity, they are subject to biases that can significantly affect downstream analysis. Bioinformatic pipelines and normalization methods reduce these biases, especially for beta-diversity calculations comparing sample community structure [34, 49].

Bioinformatic pipelines reduce bias by removing sequencing artifacts, such as single and multi-base pair variants, and chimeric sequences, from microbiome datasets. If not accounted for, these artifacts may incorrectly be attributed as novel diversity in a sample. Bioinformatic pipelines also use clustering or sequence inference techniques to group reads into biologically informative units. Standard clustering methods include *de novo* clustering based on pairwise sequence similarities [88] and closed reference clustering of reads against a reference database [29]. Open reference clustering is a combination of the two, first applying a closed reference approach, followed by *de novo* clustering of reads that did not map to a reference [82]. Sequence inference methods use statistical models and algorithms to group sequences independent of sequence similarity but based on the probability that a lower abundant sequence is an artifact originating from more highly abundant sequence, independent of sequence similarity [19, 1]. The resulting features, operational taxonomic units (OTUs) for clustering methods and sequence variants (SVs) for sequence inference methods, have different characteristics because the different methods vary in their ability to detect and remove errors while retaining true biological sequences.

Rarefaction and numeric normalization methods account for differences in sample total abundances caused by uneven pooling of samples prior to sequencing, and differences in sequencing run throughput. Rarifying abundance data traces its origins to macroecology, where counts for a unit (sample) are randomly subsampled to a user-defined constant level [35]. Although there are concerns about its statistical validity [62], rarefaction is currently the only normalization method for unweighted, presence-absence based, beta-diversity metrics [102]. For weighted, abundance based beta-diversity analyses, we can apply numeric normalization methods, such as total and cumulative sum scaling (TSS and CSS), where counts are divided by sample total abundance (TSS) or by the cumulative abundance (CSS) for a defined percentile [72]. CSS is one of the few normalization methods developed specifically for 16S rRNA marker-gene survey data. Other normalization methods, including upper quartile (UQ), trimmed mean of M values (TMM) and relative log expression [84, 58], were initially developed for normalizing RNAseq and microarray data. Many studies have found these methods useful in normalizing marker-gene survey data for differential abundance analysis, though it is unclear whether these techniques are also suitable for beta-diversity analysis.

Beta-diversity is calculated using a variety of metrics that can be grouped based on whether they account for phylogenetic distance and feature relative abundance. The UniFrac metric was developed specifically for marker-gene survey data and incorporates phylogenetic relatedness by comparing the branch lengths of features that are unique to two communities [37]. Unweighted UniFrac uses presence-absence information, whereas weighted UniFrac incorporates feature relative abundance. Taxonomic metrics do not consider the relationship between features. The Bray-Curtis and Jaccard dissimilarity indices are examples of weighted and unweighted taxonomic metrics respectively, as they do not consider the phylogenetic relationship between features [13, 44]. Because these four groups of beta-diversity metrics measure different community characteristics, they are not interchangeable should be evaluated in a complementary manner to gain maximal insight into community differences [2].

Previous studies have evaluated different bioinformatics pipelines [92] and normalization methods [62, 102] on beta-diversity analysis. Yet, the ability of these pipelines to account for sequence quality and coverage, and how this affects diversity conclusions, remains unknown. Here, we use a novel dataset of stool samples from vaccine trial participants, collected before and after exposure to the pathogen, and mixed following a two-sample titration mixture design. We sequenced multiple technical PCR replicates, allowing us to evaluate (1) beta-diversity PCR repeatability, and the ability to (2) distinguish between groups of samples with varying levels of similarity, and (3) identify differences in beta-diversity between individuals and treatment. Furthermore, the data was reproduced from across four runs with different sequencing error rates and library sizes, enabling assessment of how each pipeline and method performs on datasets of varying quality.

4.3 Methods

Our assessment framework utilizes a dataset of DNA mixtures from five vaccine trial participants described in Section 3.3.1. DNA was extracted from stool collected from five individuals (subjects) before and after exposure to pathogenic *Escherichia coli* (timepoints). The pre- and post-exposure DNA was mixed following a log_2 two-sample titration mixture design, resulting in a set of samples with varying levels of similarity. The microbial community in the unmixed pre- and post-exposure samples and titrations were measured using 16S rRNA marker-gene sequencing. Four technical replicates of each were generated during the 16S rRNA PCR amplification process. Technical replicates of each PCR were sent to two independent laboratories

(JHU and NIST) for sequencing (Fig. 4.1).

Sequencing libraries were prepared at the independent laboratories using the same protocol (16S Metagenomic Sequencing Library Preparation, posted date 11/27/2013, downloaded from `https://support.illumina.com`). Resulting libraries were sequenced twice at each laboratory, resulting in four sequence datasets with varying sequence quality and library sizes. The first JHU run PhiX error rate was higher than expected and the instrument was re-calibrated by the manufacturer, resulting in improved quality scores for the second run. The first run at NIST generated lower total throughput than expected, so the pool library for the second run was re-optimized and generated a dataset with increased throughput and lower sample to sample read count variability. No template controls were also sequenced for quality control and did not reveal any significant reagent contamination. Sequence data characterization was performed using the savR [16] and ShortRead Bioconductor R packages [63].

4.3.1 Bioinformatic Pipelines

Data from the four sequencing runs were processed using six bioinformatic pipelines, including the QIIME open reference, closed reference, *de novo*, and Deblur pipelines, as well as the Mothur *de novo* pipeline and DADA2 sequence inference pipeline. The code used to run the bioinformatic pipelines is available at `https://github.com/nate-d-olson/mgtst_pipelines/`, on the multirun branch. Preprocessing and feature detection methods vary by pipeline. The Mothur pipeline uses the OptiClust algorithm for *de novo* clustering [103]. Pre-processing includes merging and quality filtering paired-end reads followed by aligning sequences to the SILVA reference alignment [89]. Taxonomic classification was performed using the RDP Bayesian classifier [101] implemented in Mothur. The phylogenetic tree was constructed in Mothur using the clearcut algorithm [91]. Mothur version 1.39.3 (`https://www.mothur.org`) and SILVA release version 119 reference alignment and RDP the mothur formatted version of the RDP 16S rRNA database release version 10 [26].

The DADA2 big data protocol for DADA2 versions 1.4 or later was followed (`https://benjjneb.github.io/dada2/bigdata.html`), except for read length trimming parameters and primer trimming. Forward and reverse primers were trimmed using cutadapt version 1.14 (`https://cutadapt.readthedocs.io/en/stable/`) [56]. The forward and reverse reads were trimmed to 260 and 200 bp respectively. Read trimming positions were defined based on read quality score distributions, maximizing the overlap region between the forward and reverse read while minimizing the inclusion of low-quality sequence data. The pipeline was run using DADA2 version 1.6.0 [19] and formatted SILVA database version 128 trainset provided by the DADA2 developers [17]. Taxonomic classification was performed using the DADA2 implementation of the RDP Bayesian classifier [101]. The phylogenetic tree was generated following methods in [20] using the DECIPHER R package for multiple sequence alignment [108] and the phangorn R package for tree construction [87].

The QIIME pipelines all used the same merged paired-end, quality filtered set of sequences [21]. UCLUST alogrithm (version v1.2.22q) was used for clustering and taxonomic assignment against the Greengenes database version 13.8 97% similarity OTUs [29, 59]. Phylogenetic trees were constructed using FastTree, and a multiple sequence alignment generated using pyNAST and the Greengenes reference alignment [21, 78]. Both open and closed reference pipelines used the Greengenes 97% similarity database for reference clustering. Additionally, sequence variants were inferred from the QIIME merged and quality-filtered sequences using Deblur (version 1.0.3) [1]. Phylogenetic tree construction methods used for the other QIIME pipelines were also used for the Deblur pipeline.

4.3.2 Normalization Methods and Beta-Diversity Metrics

Normalization methods are used to account for between-sample differences in feature total abundance. Rarefaction, subsampling counts without replacement to an even abundance, is a commonly used normalization method in macro-ecology and 16S rRNA marker-gene surveys [35, 40]. We rarefied samples to four levels; 2000, 5000, and 10000 total reads per sample, and to the total abundance of the 15th percentile. Rarefaction levels were selected based on values used in published studies [98] and other comparison studies [102, 62]. Rarified count data were analyzed using both weighted and unweighted beta-diversity metrics. Numeric normalization methods include those previously developed for normalizing microarray and RNAseq data, such as upper quartile (UQ), trimmed mean of M values (TMM), and relative log expression [84, 58], and those that are commonly used to normalize 16S rRNA marker-gene survey, such as cumulative sum scaling (CSS) [72] and total sum scaling (proportions, TSS). Numeric normalization methods were used for weighted metrics, as they do not impact unweighted metric results.

Weighted and unweighted phylogenetic and taxonomic beta-diversity metrics were compared. Beta-diversity metrics were calculated using phyloseq version 1.22.3 [61]. Weighted and unweighted UniFrac phylogenetic beta-diversity metrics were calculated using the phyloseq implementation of FastUniFrac [61, 37]. For feature-level beta-diversity assessment, the Bray-Curtis weighted, and Jaccard unweighted metrics were used [13, 44].

4.3.3 Beta-Diversity Assessment

Standard linear models were used to test for significance using the R `lm` function. Mixed effects models, used to take into account repeated measures, were fit using the R `lmer` function in the lme4 package [6]. Model fit was evaluated based on model statistics, AIC, BIC, and logLik, as well as diagnostic plots. Tukey Honest Significant Differences test was used for multiple comparison testing using the `TukeyHSD` function. The source code for all analysis is available at `https://github.com/nate-d-olson/diversity_assessment`.

4.3.3.1 PCR Repeatability

Beta-diversity repeatability was evaluated for the different pipelines across sequencing runs. Here we define repeatability as the median beta diversity between PCR replicates. The unnormalized count data was used to characterize the baseline beta-diversity repeatability for the different pipeline and sequencing runs. Linear models were used to quantify differences between pipelines and across the four sequencing runs for the diversity metrics. Data from the first NIST sequencing run (NIST1) were used to evaluate normalization method impact on PCR replicate beta-diversity. To quantify normalization method impact, independent linear models were fit for each pipeline and diversity metric.

4.3.3.2 Signal to Noise Ratio

Next, we evaluated the signal-to-noise ratio for the different pipelines across sequencing runs by comparing pre-exposure samples to other samples in the titration series. Signal was measured as the median beta-diversity between samples were compared (Fig. 4.1). Noise was measured as the median PCR replicate beta-diversity within the compared samples. A weighted average of the signal-to-noise ratio was calculated as the area under the curve (using the `trapz` function) of the signal-to-noise ratio and the proportion of pre-exposure DNA in the sample being compared [12]. Independent linear models were fit for each diversity metric to quantify differences in the signal-to-noise ratio between sequencing runs and pipelines. A mixed-effects linear model was then used to quantify normalization method impact on the signal-to-noise ratio using data from NIST1 with subject as a random effect. Independent mixed effects linear models were fit for each pipeline and diversity metric.

4.3.3.3 Biological v. Technical Variation

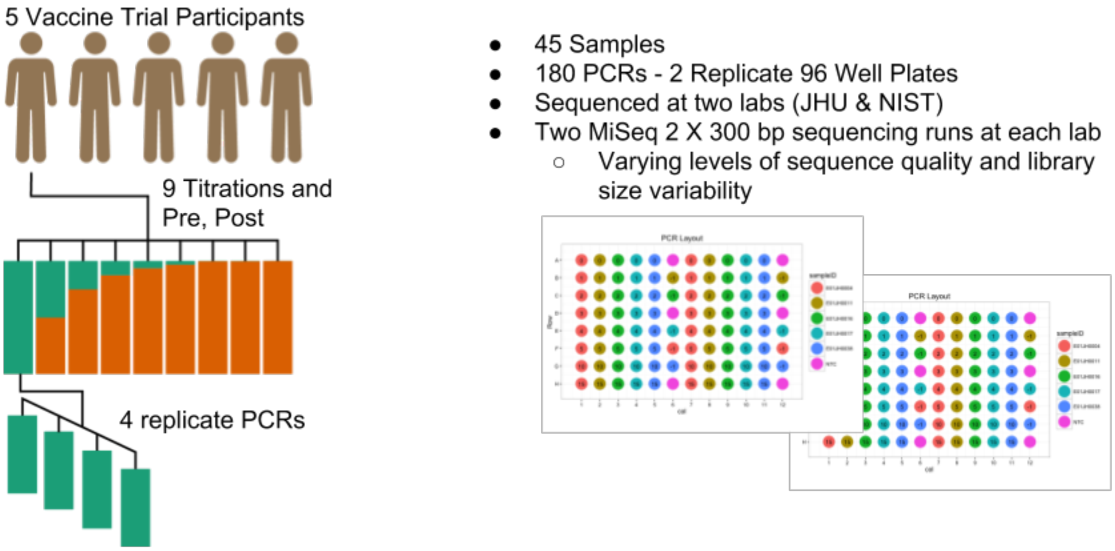
To quantify the contribution of biological and technical variability to total variability the distribution of beta diversity metrics were compared between subjects, within subject and between conditions (pre- and post-exposure), and different types of technical replicates. A linear model was used to quantify differences in beta diversity between biological and technical sources of variability. We then used variation partitioning [11] to quantify technical and biological factor's contribution to the total observed variation. Variation partition was calculated using the Vegan R package [66]. Distance-based redundancy analysis (dbRDA) was used to identify significant sources of variation [66].

4.4 Results

We sequenced the bacterial communities in stool samples collected from five vaccine trial participants before and after exposure to pathogenic *E. coli* (Fig. 4.1). Mixture samples were generated by titrating pre- and post-exposure samples at different concentrations. Each sample was sequenced twice at two different laboratories (JHU and NIST) for a total of four runs.

4.4.1 Dataset Characteristics

The four replicate sequencing runs were of variable sequence quality and depth (Fig. 4.2). Sequencing error rates and base quality scores also varied by sequencing run. JHU1 had higher PhiX error rates compared to all other runs, especially for the



- 45 Samples
- 180 PCRs - 2 Replicate 96 Well Plates
- Sequenced at two labs (JHU & NIST)
- Two MiSeq 2 X 300 bp sequencing runs at each lab
  - Varying levels of sequence quality and library size variability

Figure 4.1: Two-sample titration dataset experimental design. The dataset contained independent two-sample titration series from 5 vaccine trial participants (subjects), resulting in 45 samples. PCRs were run on two 96 well plates with each plate half containing one for each sample and three no template control reactions. The four replicate PCR assays per sample resulted in 180 PCRs. The PCR products were split into technical replicates and sequenced twice at two different laboratories.

Table 4.1: Summary statistics for the different bioinformatic pipelines. No template controls were excluded from summary statistic calculations. Sparsity is defined as the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2), rows in the count table. Singletons is the total number of features only observed once in a single sample. Total Abundance is the median and range (minimum-maximum) per sample total feature abundance. Pass Rate is the median and range for the proportion of reads not removed while processing a sample's sequence data through a bioinformatic pipeline.

Pipelines	Features	Singletons	Samples	Sparsity	Total Abundance	Pass Rate
dada	25247	99	768	0.991	52356 (141585-181)	0.76 (0.87-0.01)
mothur	38367	24490	765	0.992	13312 (42954-171)	0.2 (0.45-0.02)
q_closed	6184	829	754	0.929	24938 (111765-1)	0.36 (0.73-0)
q_deblur	3711	0	576	0.940	9135 (30423-4)	0.14 (0.24-0)
q_denovo	180834	120599	766	0.994	26250 (118767-4)	0.37 (0.75-0)
q_open	45663	39	766	0.981	26373 (118421-3)	0.37 (0.75-0)

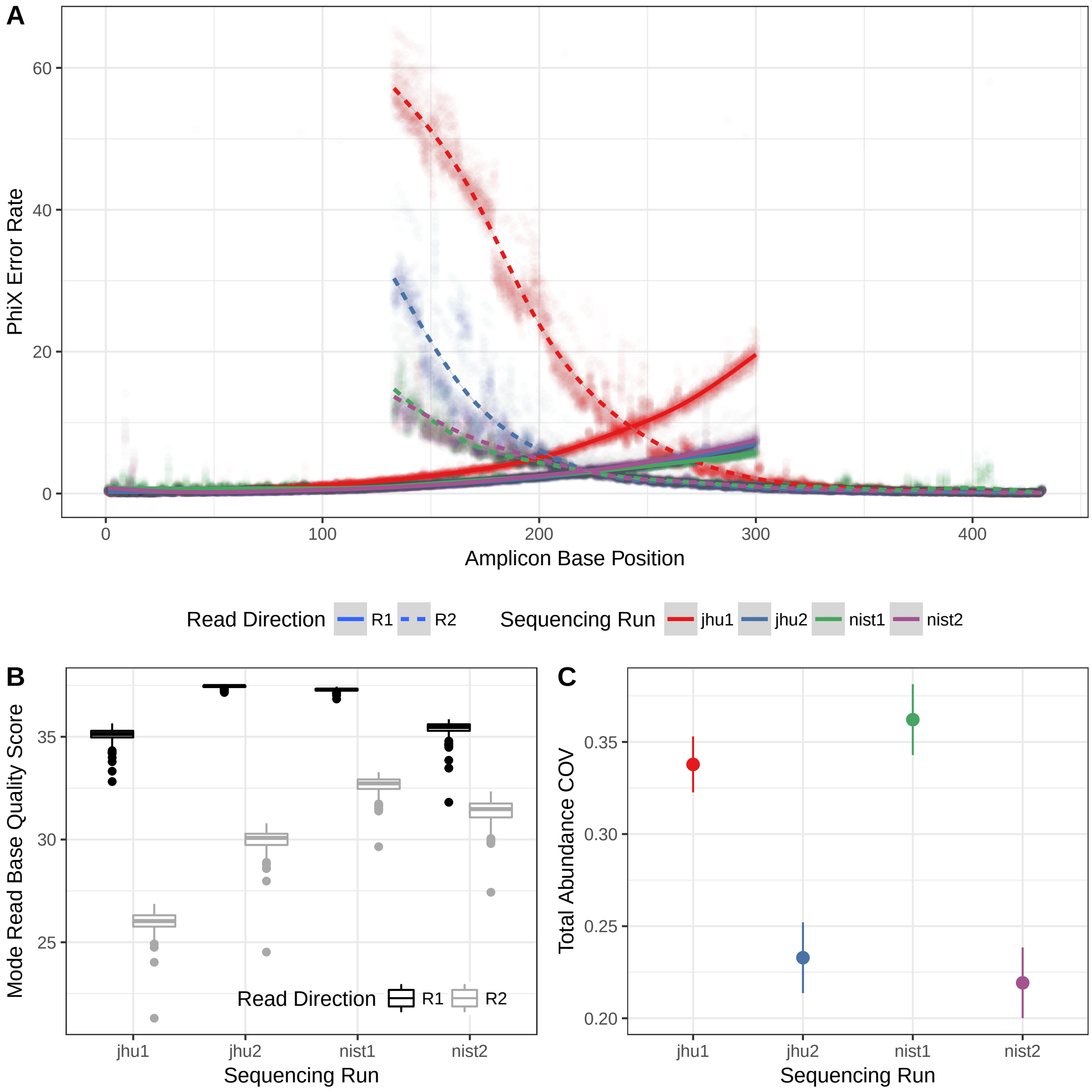


Figure 4.2: Sequencing quality and sample total abundance variation for the four sequencing runs used in this study. The same set of 192 PCRs were sequenced in all four runs. Independent sequencing libraries were generated at the two sequencing laboratories (JHU and NIST). (A) PhiX error rate relative to 16S rRNA amplicon base position for the four sequencing runs. (B) Distribution of mode read quality score by sequencing run. (C) Sequencing run total abundance coefficient of variation estimate and 95% confidence interval calculated using a mixed effects linear model.

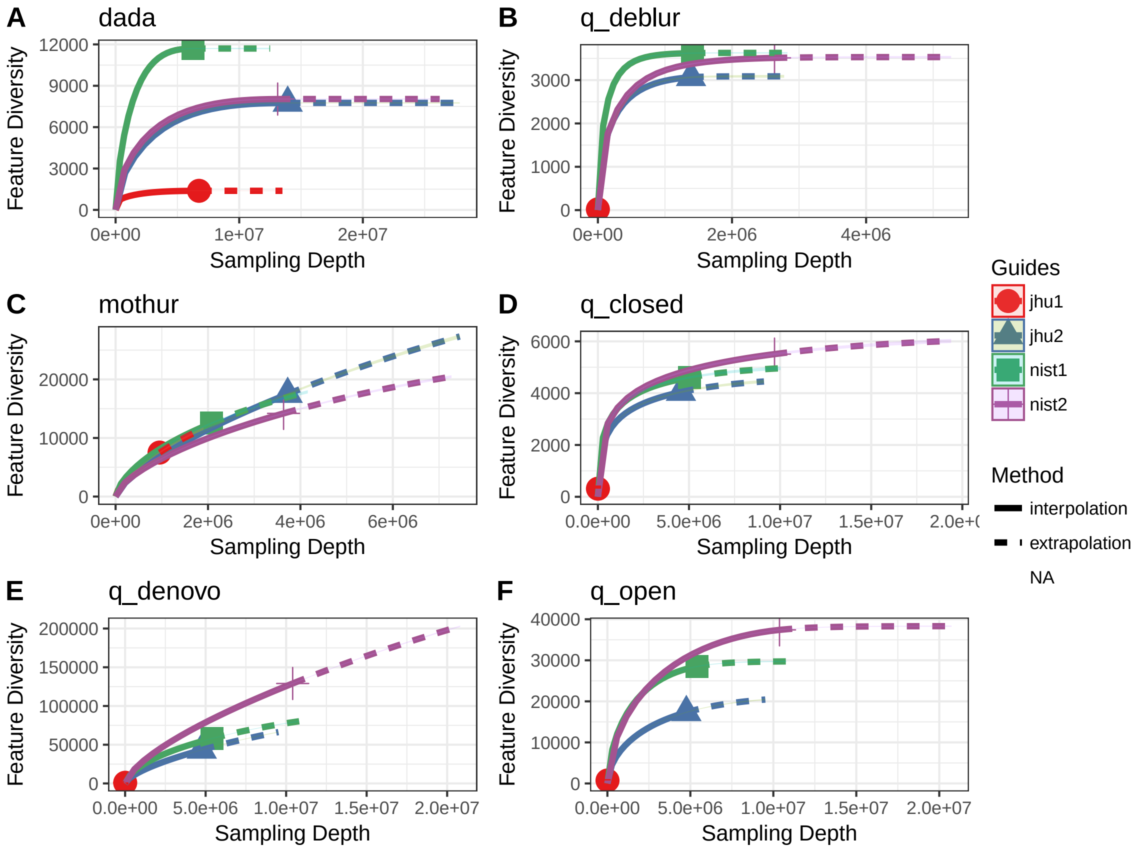


Figure 4.3: Rarefaction curves for the four sequencing runs (line color) by pipeline (A-F). Rarefaction curves were calculated using the feature counts summed across all samples by sequencing run. Rarefaction curves indicate how thoroughly a population is sampled. Curves show the relationship between the number of unique features (y-axis) and sampling depth. Curves reaching an asymptote indicate the population has been completely sampled. Shapes indicate the observed feature diversity and sampling depth. Solid lines represent interpolated values obtained by randomly subsampling the observed abundance data. Dashed lines indicate extrapolated values predicted based on the observed count data and interpolated values.

reverse reads (Fig. 4.2A). Read base quality was lower for the reverse read than the forward reads for all four sequencing runs (Fig. 4.2B). Sequence data from the two NIST runs had higher quality scores than the data from JHU runs, except for JHU2 forward reads (Fig. 4.2B). Greater variability in sample feature total abundance was observed on the first run at each laboratory (Fig. 4.2C).

Overall, sequences from JHU1 had lower read quality and higher variability in total sample abundance. Sequences from NIST1 were of higher quality but also exhibited greater variability in total sample abundance. Thus, by comparing the JHU1 results to the higher quality, less variable NIST2 and JHU2 runs, we can evaluate how well the bioinformatic pipelines handle low quality reads. Similarly, we can use data from the NIST1 to determine how well normalization methods can account for differences in total abundance between samples.

Samples from the different sequencing runs were processed using six different bioinformatic pipelines. Four of the pipelines, including the QIIME *de novo*, QIIME closed-reference, QIIME open-reference, Mothur *de novo*, utilize OTU clustering methods, while the remaining two, QIIME Deblur and DADA2, use sequence inference approaches. Aside from the four QIIME pipelines each pipeline employs its own pre-processing, feature inference, and quality filtering methods. The four QIIME pipelines used the same pre-processing methods. As a result, the features and count tables generated by the pipelines exhibit different characteristics in terms of the number of features, total abundance, number of singletons, the proportion of sequences passing quality control (Table 4.1).

We generated rarefaction curves to assess feature diversity at multiple sampling depths for across the four sequencing runs (Fig. 4.3). Sequence inference methods (DADA2 and Deblur) had lower overall feature diversity estimates and their rarefaction curves reached an asymptote around the same level (Fig. 4.3A & B), suggesting that sampling depth was sufficient to capture community diversity. The JHU1 rarefaction curves at the origin for the QIIME pipelines was due to limited number of features, none for Deblur, were produced by the pipelines. DADA2 asymptotes, however, were inconsistent across sequencing runs, indicating artificial plateaus for the lower throughput and lower quality runs (Fig. 4.3A). Rarefaction curves for *de novo*, open-reference, and closed-reference methods did not reach an asymptote (Fig. 4.3). The QIIME *de novo* pipeline had the greatest slope, suggesting the highest rate of artifacts (Fig. 4.3E). This is most likely due to the fact that the QIIME *de novo* pipeline does not filter out singletons (Table 4.1). Furthermore, the Mothur rarefaction curves were consistent across sequencing runs, but the QIIME clustering pipelines rarefaction curves were influenced by both sequence quality and library size (Fig. 4.3D-F).

4.4.2 PCR Repeatability

Next, we evaluated differences in beta-diversity between un-normalized PCR replicates across sequencing runs and pipelines. PCR replicate beta-diversity varied by diversity metric (Fig. 4.2). Beta-diversity was consistently higher for unweighted compared to weighted metrics, and phylogenetic diversity metrics were lower than taxonomic metrics. We expected to see higher pairwise distances for the lower

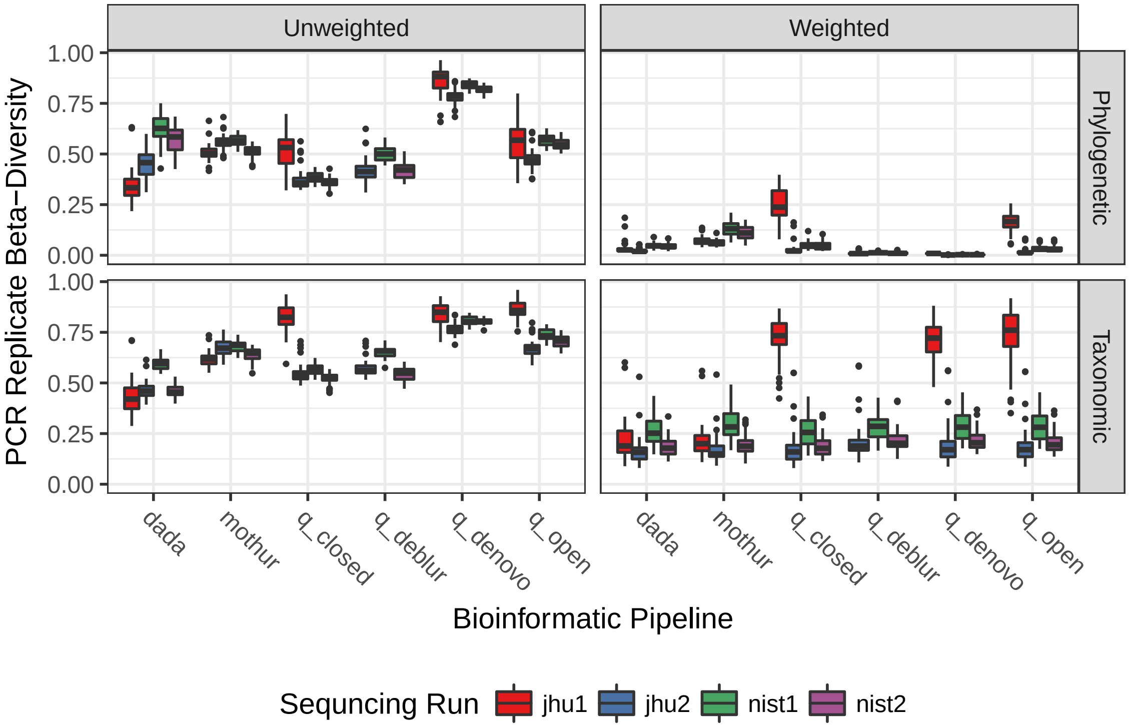


Figure 4.4: Distribution of mean pairwise PCR replicate beta-diversity by sequencing run and pipeline for un-normalized count data.

quality JHU1 run compared to the higher quality JHU2 run. This was true for the QIIME clustering pipelines. However the Mothur and DADA2 mean PCR replicate beta-diversity was consistent across the JHU runs, suggesting that these pipelines are more robust to sequencing errors (Fig. 4.2). Conversely, with the highest number of failed samples for the first JHU run, the Deblur pipeline was the least robust to sequencing errors (Table 4.1). As expected JHU2 and NIST2, with high read quality and lower total abundance variability, had comparable PCR replicates beta-diversity. Additionlly, NIST1 had higher PCR replicate beta-diversity compared to JHU2 and NIST2, which is attributed to higher total abundance vairiability.

Data from NIST1 was used to compare normalization methods ability to improve beta-diversity repeatability. When comparing normalized to un-normalized

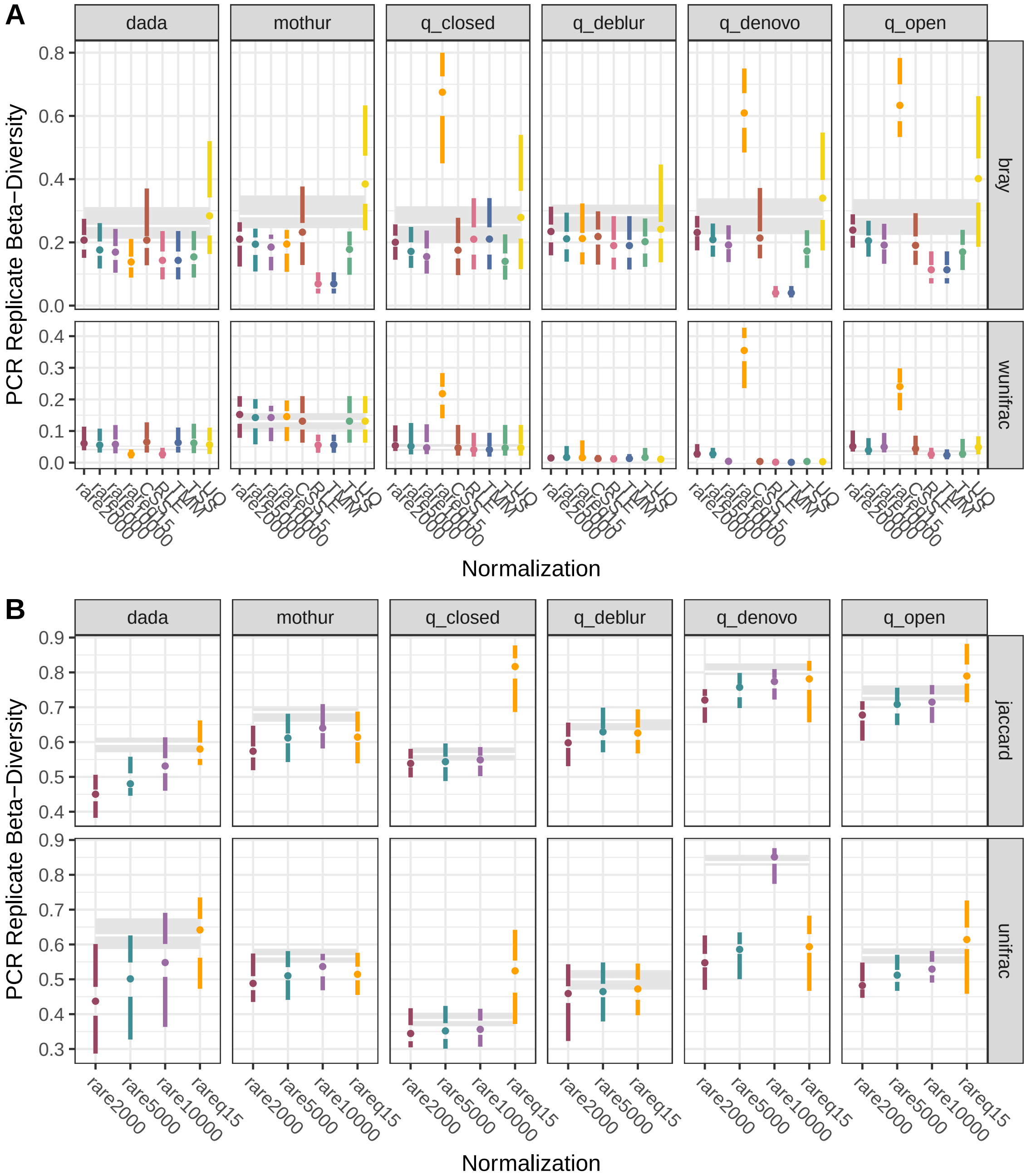


Figure 4.5: Impact of normalization method on mean weighted (A) and unweighted (B) PCR replicates beta-diversity, for the sequencing run with higher quality and total abundance variability, NIST1. Data are presented as minimal-ink boxplots, where points indicate median value, the gap between point and lines the interquartile range, and lines the boxplot whiskers. Solid black lines represent median value and dashed lines indicate the first and third quartiles of the raw (un-normalized) mean pairwise distances between PCR replicates.

PCR replicate beta-diversity, we observed that most normalization methods reduced beta-diversity between PCR replicates (Fig. 4.5A). For a number of pipelines, TMM and RLE normalization methods significantly lowered weighted PCR replicate beta-diversity (Fig. 4.5A). For unweighted metrics (Fig. 4.5B), rarefying count data to 2000 total feature abundance resulted in the lowest beta-diversity between PCR replicates. While rarefying counts to the total abundance of the 15th most abundant sample (rareq15) tended to significantly increase PCR replciates beta-diversity. Rarefaction to this level is also most susceptible to sample loss and should not be used as it results in unnecessary loss of statistical power.

4.4.3 Signal to Noise

We further sought to identify which pipelines and normalization methods are best able to pull out biological signals from background, technical noise. We calculated a signal-to-noise ratio by dividing the beta-diversity between unmixed pre-exposure samples and other samples in the titration series (signal) by PCR replicate beta-diversity for the samples being compared. The signal-to-noise ratio for unweighted metrics on un-normalized samples was around 1 for all pipelines and sequencing runs (Fig. 4.6), indicating that the signal magnitude (biological differences) was equal to the noise (differences between PCR replicates). Using weighted metrics, only DADA2 and Mothur ratios were consistently greater than 1, and these pipelines had higher ratio differences for the JHU runs compared to NIST runs. The relationship between NIST and JHU runs for the signal to noise relationship is

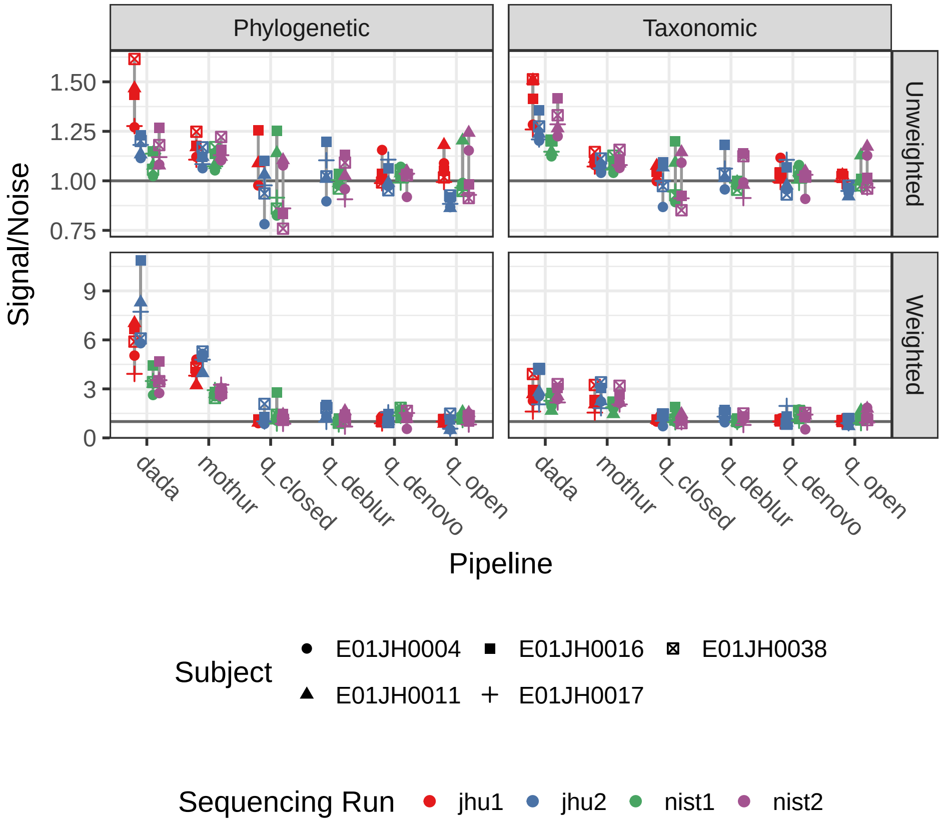


Figure 4.6: The weighted average signal to noise varied by pipeline, run, and diversity metric. Points indicate the signal to noise for each individual with grey lines representing the range of values for a pipeline and sequencing run. Dark grey horizontal lines indicate a signal-to-noise ratio of 1.

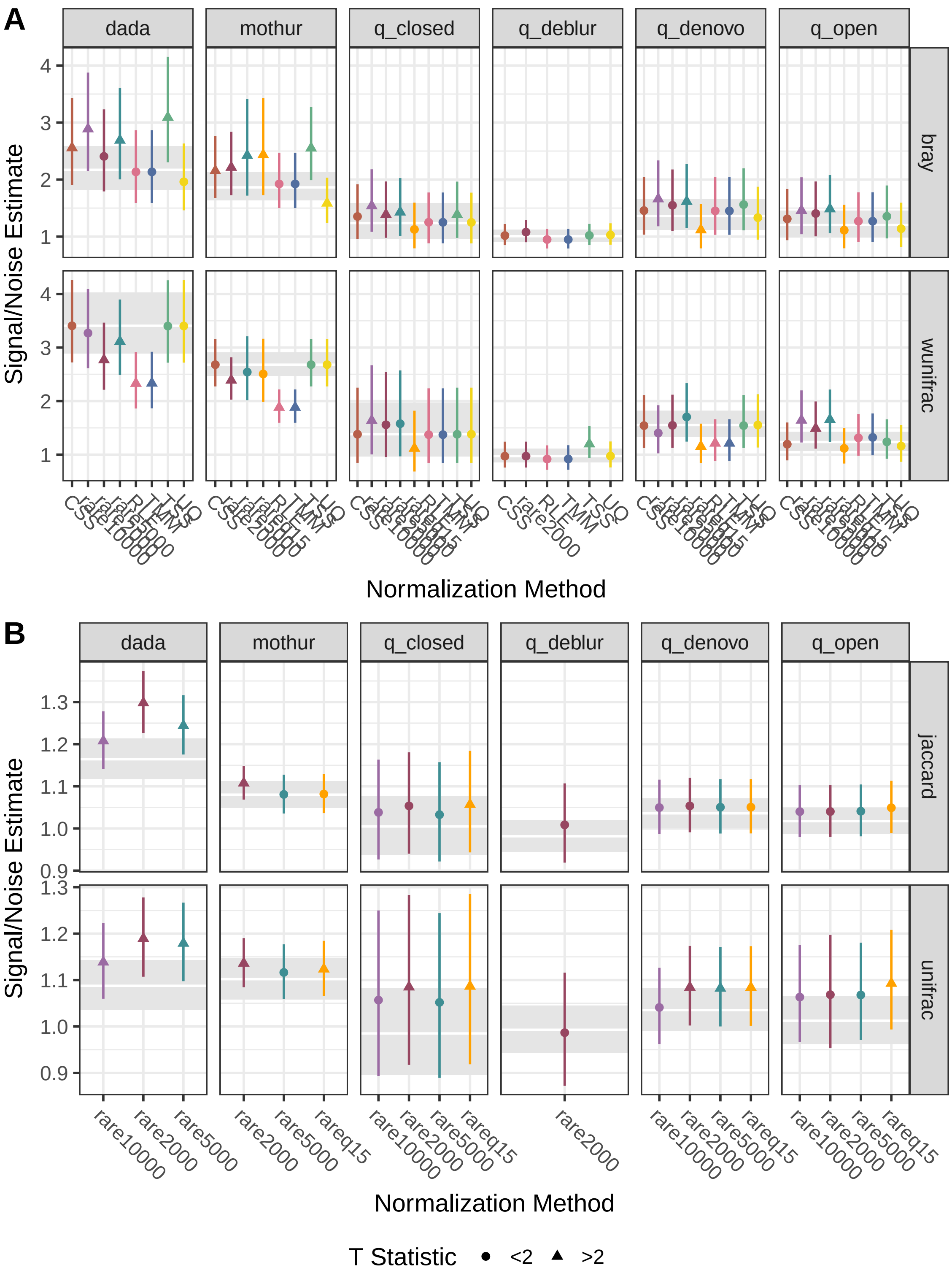


Figure 4.7: Weighted average signal to noise ratio estimate and 95 CI for raw and normalized count data for (A) weighted and (B) unweighted beta-diversity metrics. Estimates were calculated using a mixed effects linear model using subject as random effect. The horizontal solid line is the unnormalized count signal to noise estimate,S and horizontal dashed lines indicate 95 CI. The points and line ranges indicate the model estimate and 95 CI for the different normalization methods.

consistent with the PCR replicate beta-diversity results.

Normalizing count data should increase the signal-to-noise ratio; however, most normalization methods did not have a significant for weighted metrics (Fig. 4.7A). One exception was TSS, which significantly increased the Bray-Curtis signal to noise ratio for the Mothur and DADA2 datasets. Rarefying counts to the 15th quantile resulted in significantly lower the weighted UniFrac and Bray-Curtis signal-to-noise ratio for QIIME closed-reference and *de novo* pipelines. While RLE and TMM improved PCR replicate beta-diversity, these normalization methods also significantly lowered the weighted UniFrac beta-diversity for DADA2, Mothur, and QIIME *de novo* pipelines. Rarefaction often increased the unweighted metric signal-to-noise ratio (Fig. 4.7B), though the increase was only significant at lower subsampling depths for DADA2 and Mothur pipelines.

4.4.4 Biological v. Technical Variation

Finally, we characterized how different pipelines and normalization methods capture diversity differences between biological factors and technical replicates. As expected, the mean diversity observed between biological factors was greater than between technical replicates (Fig. 4.8). The magnitude of this difference, however, was greater for weighted than unweighted beta-diversity metrics and varied by pipeline. Greater differences were observed with the DADA2, Mothur, and Deblur pipelines, compared to the QIIME clustering approaches.

Variation partitioning was used to identify the amount of variation attributable

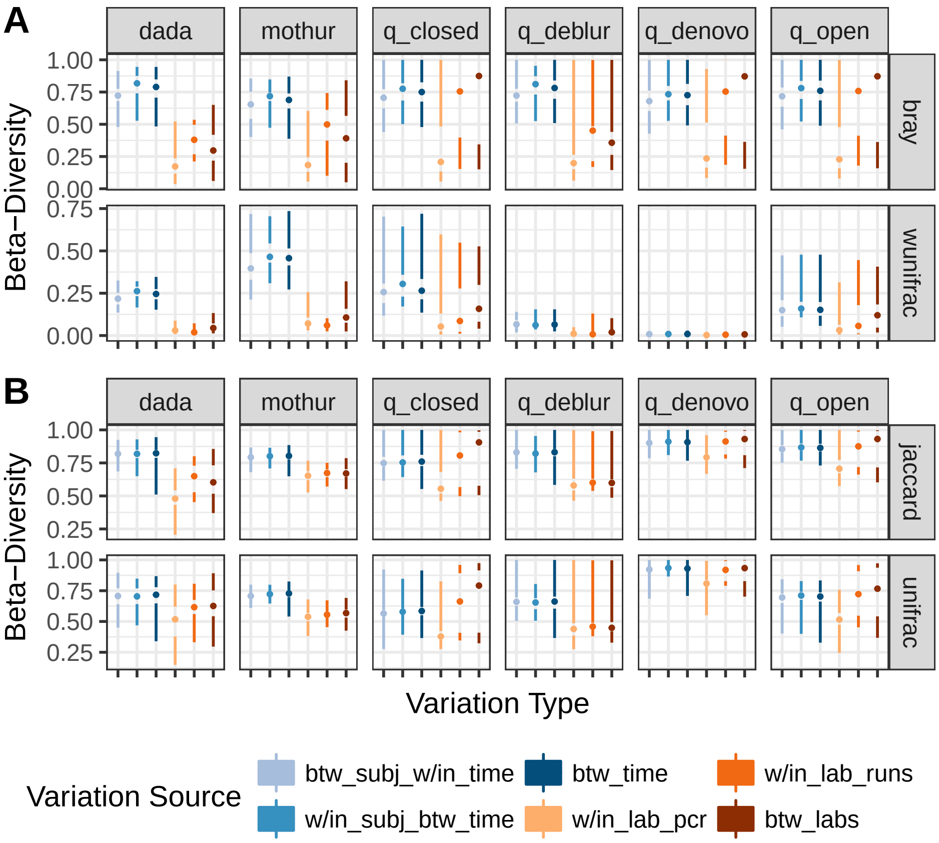


Figure 4.8: Biological vs. Technical Variation, distribution in (A) weighted and (B) unweighted beta-diversity between technical replicates and biological treatments (subject and timepoint).

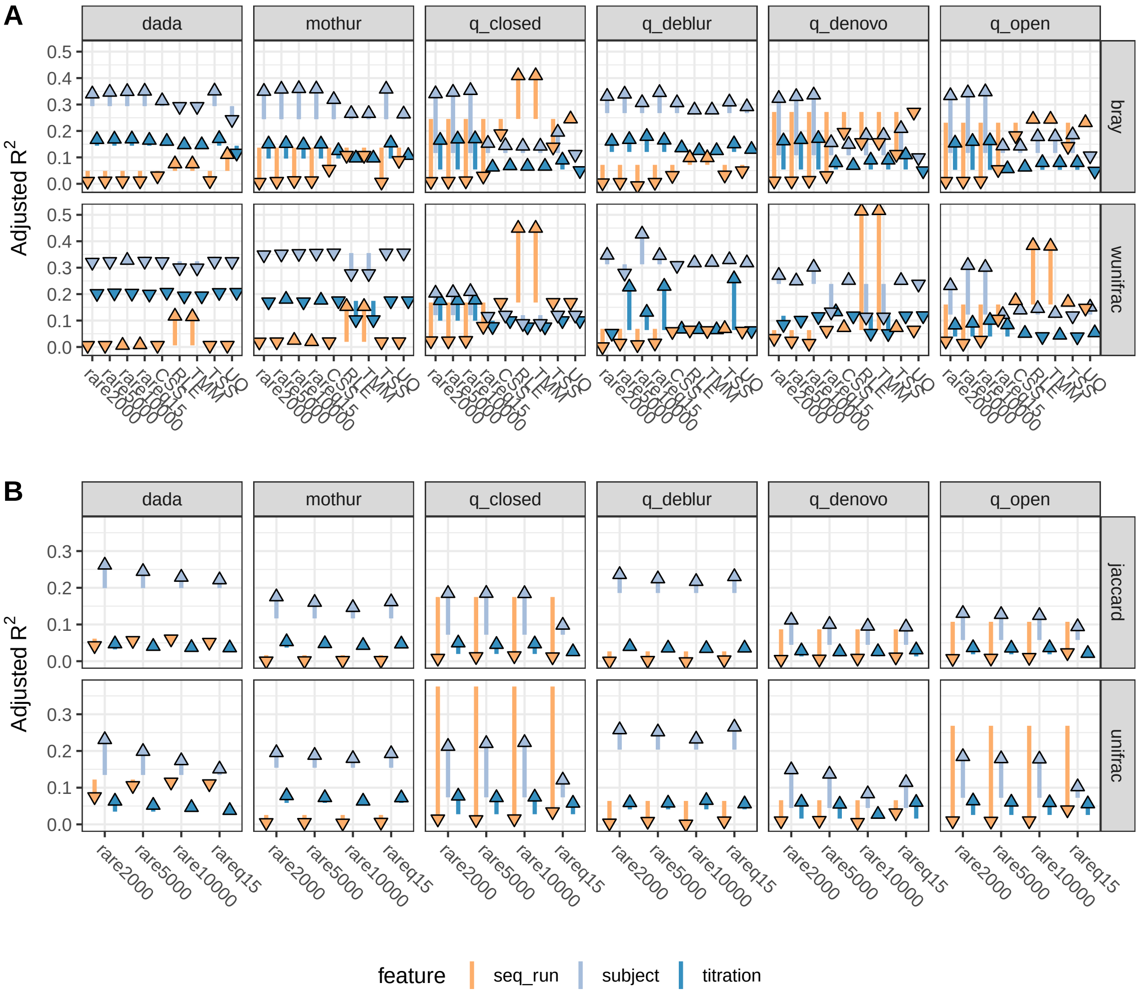


Figure 4.9: Impact of different normalization methods on biological and technical sources of variatoin for different pipelines and beta-diversity metrics. y-axis is the adjusted R^2, indicating the proportion of variance explained by each biological (subject and titration) and technical (seq run) variable. Normalized adjusted R^2 values greater than and less than unnormalized values indicated with upright triangle and upsidedown triangles, respectively. Vertical lines indicate difference between unnormalized and normalized adjusted R^2 values.

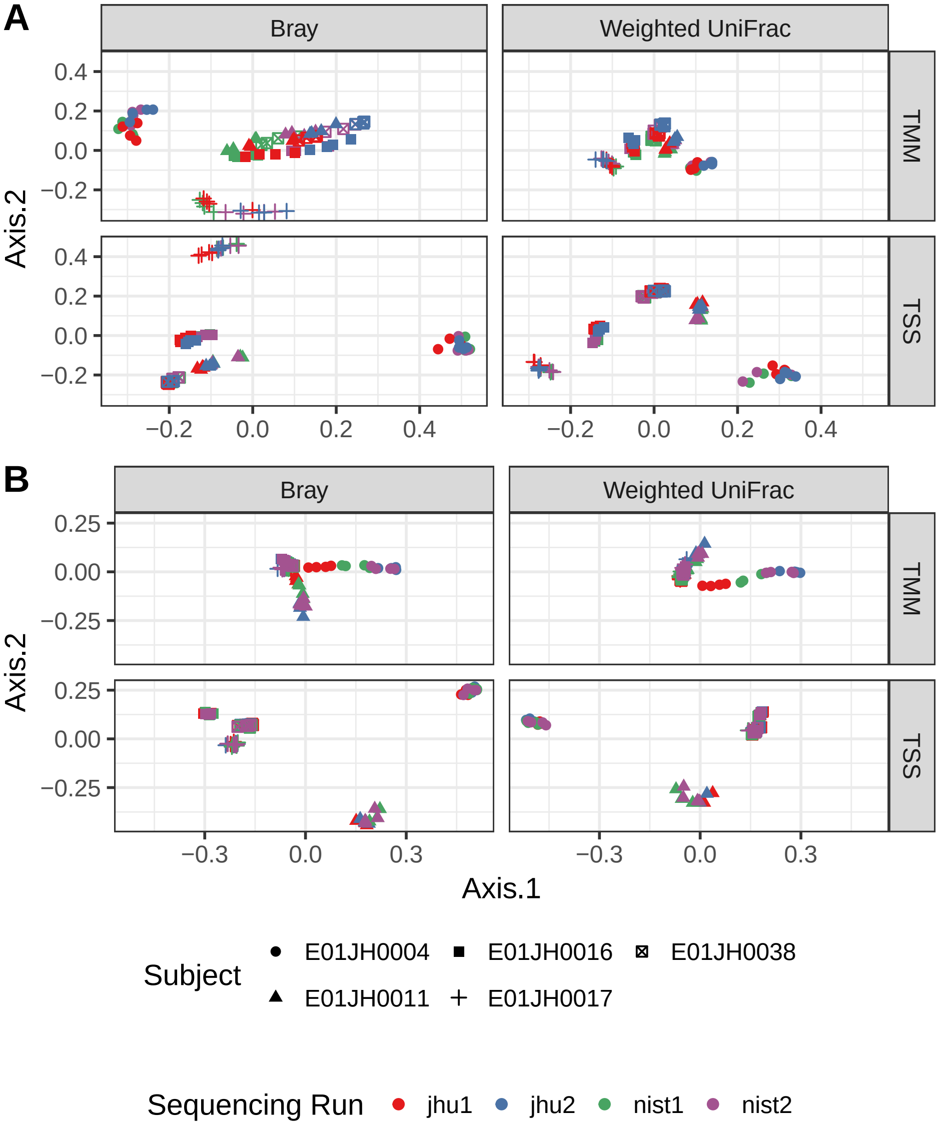


Figure 4.10: Principal coordinate analysis for TMM and TSS normalized (A) DADA2 and (B) Mothur unmixed PRE samples for Bray-Curtis and Weighted UniFrac distance metrics.

to subject, titration factor (unmixed pre-exposure and unmixed post-exposure), and sequencing run. When a normalization method increases the variation in the data (distance matrix) for a biological factor and decreases the variation for a technical factor, the beta-diversity between biological samples (i.e. different subjects) increases and beta-diversity between technical replicates (i.e. PCR assays) decreases. When beta-diversity between biological factors is equivalent to or smaller than beta-diversity between technical factors the method is no longer able to distinguish between the biological samples. Therefore the expectation is that normalization methods should decrease variation attributed to technical factors with either no change or increase the variation due to biological factors. Across all pipelines and diversity metrics, the greatest amount of variation is often explained by subject, followed by titration factor (Fig. 4.9). The variation partitioning results are consistent with our observation of greater biological than technical variability. Sequencing run accounts for a greater proportion of the explained variance in the unnormalized runs, highlighting the overall importance of normalizing our datasets.

Effective normalization methods decrease technical noise in the data without decreasing biological signal. For both weighted (Fig. 4.9A) and unweighted (Fig. 4.9B) metrics, rarefaction normalization methods show increased proportion of variation explained by biological factors and decreased the proportion of variation explained by technical artifacts. Numeric normalization methods were not as effective, especially for the QIIME pipelines. RLE and TMM normalization consistently increased technical variability and often decreased biological variability (Fig. 4.9A). Principal coordiante analysis plots for the unmixed pre-exposure samples are

Table 4.2: Pipeline beta-diversity assessment summary. +/- were used to qualitatively summarise performance of the six pipelines in for the three assessments.

Pipelines	PCR Repeatability	Signal-to-Noise	Biological v. Technical
dada	+	+	+
mothur	+	+	+
q_closed	+	-	-
q_deblur	+	-	+
q_denovo	-	-	-
q_open	-	-	-

consistent with variation partitioning results (Fig. 4.10). For Mothur and DADA2 the technical replicates group more tightly when TSS is used to normalize count data compared to when TMM.

4.5 Discussion

Sequence error rate and variation in library size are just two sequencing characteristics that can negatively bias beta-diversity analyses [62]. Ideally, bioinformatic pipelines can help differentiate true biological sequences from artifacts generated by sequencing errors [19] and normalization methods, such as rarefaction and total sum scaling, can adjust for differences in library size [72]. However, the efficacy of these different pipelines and normalization techniques for microbiome datasets, and how they affect study conclusions, are not well characterized. We compared the performance of six bioinformatic pipelines and nine normalization methods on mixture samples for four beta-diversity metrics, finding that these pipelines and methods vary significantly in their ability to identify and correct these biases.

We utilized a novel two-sample titration dataset of DNA extracts from five participants in a vaccine trial. Individual titration series were generated for each participant, where DNA collected before exposure to pathogenic *E. coli* were titrated into DNA samples collected after exposure. These samples were processed with multiple levels of technical replication, including 16S rRNA PCR assays, sequencing libraries, and sequencing runs that were performed in duplicate at two independent laboratories. Our framework assessed three components: (1) beta-diversity repeatability of PCR replicates, (2) signal-to-noise analysis of the between to within-sample beta diversity of titration sets, and (3) contribution of biological (subjects and exposure status) and technical factors (PCR replicates, sequencing labs, and runs) to beta-diversity. Pipeline performance for the three assessments are summarized in Table 4.2.

When comparing PCR replicates for all sequencing runs, the QIIME *de novo* pipeline had high UniFrac values, but low weighted UniFrac values. This is most likely due to the high proportion of singletons generated (Table 4.1). A large number of singletons indicates that a pipeline is unable to group sequencing artifacts with true biological sequences. Beta-diversity measures the relationship between single sample diversity (alpha) and system diversity (gamma). Inflated alpha- and gamma-diversity due to spurious features, as observed with QIIME *de novo* will result in inflated beta-diversity, and spurious features have a low probablility of being observed in both samples. The removal of singletons, a step included in many workflows such as the QIIME open-reference pipeline, can address this bias. Deflated alpha- and gamma-diversity, as observed with DADA2, due to grouping low abundance features with high abundance features, can similarity result in inflated beta-diversity when shared features are incorrectly grouped with non-shared features. The differences we observed in weighted and unweighted Unifrac values also emphasize the importance of assessing multiple beta-diversity metrics, as each metric provides unique insight into community composition shifts. Normalization methods generally improved beta-diversity repeatability, with the exception of rarefying data to 15th quantile, which resulted in higher beta-diversity between PCR replicates, especially for QIIME pipelines, possibly due to large sample loss. Count data normalized using TMM and RLE consistently had lower beta-diversity values between PCR replicates compared to un-normalized count data.

The biological signal magnitude was equal to the technical noise for un-normalized samples, highlighting the overall importance of normalization. Rarefaction methods at lower subsampling depths generally increased the signal to noise ratio for unweighted metrics, especially for the DADA2 and Mothur pipelines. Unexpectedly, most numeric normalization methods did not increase the signal-to-noise ratio for weighted metrics, and TMM and RLE normalization methods, which showed the greatest similarity between PCR replicates, decreased our ability to tease out the true biological indicators.

We finally evaluated the impact of different sources of variability on pipeline and normalization methods by comparing diversity between biological samples and technical replicates. For most pipelines and beta diversity metrics, normalizing the count data increased the difference in beta diversity between biological and technical replicates (Fig. 4.5), indicating a greater ability to detect community levels differences between treatment conditions. Some metrics, namely rarefying to the 15th quantile, RLE, and TMM, frequently reduced the differences in beta-diversity between the biological to technical factors. Variation partitioning results were consistent with this conclusion (Fig. 4.9).

This study highlights the importance of rigorous evaluation of computational tools and datasets. While we utilized six commonly cited bioinformatics pipelines, there are many different approaches and researchers should think critically about which is most appropriate for their own dataset. We used default program parameters in our analyses to make our findings generally applicable. However, we strongly advise researchers to have a good understanding of each step in their chosen pipeline, including what parameters are required and whether they should be changed to best fit data of interest.

Furthermore, this study shows the importance of normalizing microbiome count tables prior to beta-diversity analyses. As the microbiome field is relatively young, many existing normalization approaches are adopted from methods created for other applications. For instance, RLE and TMM normalization methods were initially developed for normalizing microarray and RNAseq data, not marker-gene sequence data. While these methods improve differential abundance analysis [62], they may not appropriate for beta-diversity analysis.

4.6 Conclusions

The results presented in this study can be used to help determine appropriate bioinformatic pipeline and normalization method for a marker-gene survey beta-diversity analysis. The six pipelines evaluated in this study varied in their ability to distinguish sequencing artifacts from true biological sequences and these differences impacted the PCR replicate beta-diversity repeatability. Based on our study results we found Mothur and DADA2 to be more robust to lower quality sequence datasets. Optimizing QIIME preprocessing methods may increase pipeline robustness to lower quality data. Additionally, the assessment presented here evaluated full bioinformatic pipelines, including both pre-processing and feature inference methods. Using the same set of pre-processed sequence data would allow for an independent evaluation of the feature inference methods. Overall, we recommend using Mothur when processing 16S rRNA sequencing data for beta-diversity analysis. Mothur was more robust to low-quality sequence data, had consistent rarefaction curves between sequencing runs, and performed well in our assessment. Additionally, as 24,490 of the 38,367 Mothur features were singletons, singleton removal will likely improve the assessment results.

Normalization can improve PCR replicate repeatability, but sometimes at the cost of decreasing the differences in beta-diversity for biological relative to technical factors. Our results indicate normalization methods developed for gene expression data analysis may not be appropriate for marker-gene survey beta-diversity analysis. For weighted metrics, we recommend normalizing counts using TSS and CSS. These normalization methods improved assessment results or had no effect relative to unnormalized counts. Rarefying count data improved unweighted metric results but higher rarefaction levels tended to perform worse than unnormalized data. Rarefying counts lowers statistical power and therefore, it is not advisable when other normalization methods are available [62]. As numeric normalization methods are not applicable to unweighted metrics, rarefying counts is the recommended normalization method. To reduce the risk of the random subsampling step biasing beta diversity results bootstrap replicates can be used to validate results.

Bioinformatic pipelines combine multiple algorithms to convert raw sequence data into a count table which is subsequently used to test biological hypotheses. Algorithm choice and parameters can significantly impact pipeline results. The pipelines compared in this study were optimized using mock communities and benchmarked against other methods based on similarity in beta-diversity results [10]. The novel assessment framework and dataset presented here provides complementary methods for use in optimizing existing and benchmarking new pipelines and normalization methods.

CHAPTER 5

metagenomeFeatures

An R package for working with 16S rRNA reference databases and marker-gene survey feature data.

5.1 Abstract

We developed the metagenomeFeatures R Bioconductor package along with annotation packages for the three primary 16S rRNA databases (Greengenes, RDP, and SILVA) to facilitate working with 16S rRNA sequence databases and marker-gene survey feature data. The metagenomeFeatures package defines two classes, `MgDb` for working with 16S rRNA sequence databases, and `mgFeatures` for working with marker-gene survey feature data. The associated annotation packages provide a consistent interface to the different 16S rRNA databases facilitating database comparison and exploration. The `mgFeatures` represents a crucial step in the development of a common data structure for working with 16S marker-gene survey data in R.

Availability: `https://bioconductor.org/packages/release/bioc/html/me`

`tagenomeFeatures.html`

5.2 Introduction

16S rRNA marker-gene surveys have significantly advanced our understanding of the diversity and structure of prokaryotic communities present in ecosystems including the human gut, open ocean, and even the international space station [53, 98, 41]. For a 16S rRNA marker-gene survey, the 16S rRNA gene is sequenced using a targeted assay. The raw sequence data is processed using a bioinformatic pipeline where the sequences are grouped into features, e.g., operational taxonomic units (OTUs) or sequence variants (SVs), yielding a set of representative sequences [18, 7].

A critical step in 16S rRNA marker-gene surveys is comparing representative sequences to a reference database for taxonomic classification or phylogenetic placement [65]. There are numerous 16S rRNA reference databases of which Greengenes, RDP, and SILVA are arguably the most commonly used [28, 26, 79, 59]. Additionally, there are smaller system-specific databases such as HOMD for the human oral microbiome [22, `http://www.homd.org/`] and soil reference database [24]. System-specific databases can improve taxonomic assignments for microbial communities not well represented in the major databases [85].

16S rRNA databases differ in the number and diversity of sequences, the taxonomic classification system, and the inclusion of intermediate ranks [5, Table 5.1]. Databases format their data differently and use sequence identification systems

Table 5.1: 16S rRNA gene sequence databases with Bioconductor annotation packages we developed.

Database	Version	Sequences	Taxonomic System
Greengenes	13.5	1,262,986	NCBI
RDP	11.5	3,356,809	Bergeys
SILVA	128.1	1,922,213	Bergeys

unique to their database, challenging membership and composition comparisons. For example, Yang, Wang, and Qian [110] used the SILVA database to evaluate how different 16S rRNA variable regions impact phylogenetic analysis. Similarly, Martinez-Porchas et al. [57] also used the SILVA database to evaluate sequence similarity between 16S rRNA gene conserved regions. Differences in database formatting present a significant barrier to performing the same analysis using multiple databases. Additionally, taxonomic assignments can be database-dependent, providing further justification for database comparisons [75]. To facilitate database comparisons RNACentral (`http://rnacentral.org/`) a resource combining non-coding RNA databases, provides unique identifiers for the sequences [96].

We developed the R package *metagenomeFeatures* for working with both 16S rRNA gene database and marker-gene survey feature data. *metagenomeFeatures* provides a common data structure for working with the 16S rRNA databases and marker-gene survey feature data. Additionally, this package is the first step towards the development of a common data structure for use in analyzing metagenomic and marker-gene survey data using R packages such as *phyloseq* [60] and *metagenomeSeq* [73].

5.3 MgDb and mgFeatures Class Description

The *metagenomeFeatures* package defines two data structures , MgDb for working with 16S rRNA databases, and mgFeatures for working with marker-gene survey feature data. There are three types of relevant information for both MgDb and mgFeatures class objects, (1) the sequences themselves, (2) sequence taxonomic lineage, and (3) a phylogenetic tree representing the evolutionary relationship between features. MgDb and mgFeatures data structures are both S4 object-oriented classes with slots for taxonomic, sequence, phylogenetic, and metadata.

As the 16S rRNA databases contain hundreds of thousands to millions of sequences, an SQLite database is used to store the taxonomic and sequence data. Using an SQLite database prevents the user from loading the full database into memory. The database connection is managed using the *RSQLite* R package [64]). The taxonomic data are accessed using the *dplyr* and *dbplyr* packages [105, 107]. The *DECIPHER* package is used to format the sequence data as an SQLite database [108]. The phylo class, defined in a *APE* R package, is used to define the tree slot [70]. We developed Bioconductor annotation packages for commonly used databases, Greengenes, RDP, and SILVA Table [26, 79, 28, Table 5.1]. Along with database specific sequence identifiers, RNAcentral identifiers are included in the SQLite table for inter-database comparisons.

mgFeatures-class is used for storing and working with marker-gene survey feature data. Similar to the MgDb-class, the mgFeatures-class has four slots, for taxonomy, sequences, phylogenetic tree, and metadata. As the number of features in a marker-gene survey dataset is significantly smaller than the number of sequences in a reference database, `mgFeatures-class` uses common Bioconductor data structures, `DataFrames` and `DNAStringSets` to define the taxonomic and sequence slots [69, 68]. Similar to `MgDb-class`, a `phylo-class` object is used to define the tree slot. For both the `MgDb` and `mgFeatures` classes the tree slot is optional, and the metadata are stored as a list.

5.4 Applications/ Vignettes

The metagenomeFeatures package includes a series of vignettes as example use cases for the metagenomeFeatures package and associated reference database annotation packages. (1) Retrieving sequence and phylogenetic data for OTUs from closed-reference clustering. (2) Exploring the diversity of a taxonomic group of interest.

The R command browseVignettes("metagenomeFeatures") provides a list of vignettes associated with the package and vignette("x") is used to view specific vignettes, where "x" is the vignette name.

To further demonstrate the utility of the package, the manuscript supplemental information uses *metagenomeFeatures*, *greengenes13.5MgDb* annotation package, and *DECIPHER* to evaluate the potential for species-level taxonomic classification using 16S rRNA V12 and V4 sequence data.

5.5 Conclusions

The *metagenomeFeatures* package provides data structures and functions for working with 16S rRNA databases and marker-gene survey feature data. The data structure provided by the `MgDb-class` in conjunction with the shared sequence identifier system developed by RNACentral facilitates comparisons between 16S rRNA databases. The `mgFeatures-class` provides the groundwork for the development of a common data structure for working with metagenomic and marker-gene sequence data in R which will increase interoperability between R packages developed for working with metagenomic sequence data. Additionally, while the data structures were developed for 16S rRNA gene sequence data they can be used for any marker-gene sequence data without modification and can be extended to work with shotgun metagenomic sequence data and databases.

5.6 Supplemental Material

Paenibacillus species resolution for 16S rRNA V12 and V4 regions.

5.6.1 Background

16S rRNA amplicon sequencing is commonly used for microbial community characterization, including differential abundance analysis. A limitation to 16S rRNA amplicon sequencing is a lack of taxonomic resolution, where organisms are only identifiable to the genus or family level. We define taxonomic resolution as the ability to differentiate between groups within a taxonomic level, for example differentiating between species within a genus. While similar to determining whether a sequence represents a novel species, here we are only interested in determining whether the 16S rRNA region of interest contains sufficient information for species-level taxonomic assignment. Taxonomic resolution varies by clade and amplicon regions. Though the extent to which taxonomic resolition varies is not well characterized.

Here we demonstrate how *metagenomeFeatures* and the `MgDb` annotation packages can be used to characterize taxonomic resolution for a specific clade and amplicon region, specifically for the *Paenibacillus* genus and V12 and V4 regions. Originally classified under the *Bacillus* genus, a novel genus was formed based on the 16S rRNA gene similarity in the 1990s. *Paenibacillus* spp. are facultative anaerobic bacteria present in a variety of environments including the soil, water, and can act as opportunistic pathogens in humans [67]. *Paenibacillus* spp. will play an important role in sustainable agricultural industries [36]. As such, appropriate speciation is of interest. The V12 and V4 region were used as they represent two commonly used amplicons for 16S rRNA marker-gene surveys. We will use the Greengenes 13.5 database, accessed using the *greengenes13.5MgDb* annotation package for our analysis of the *Paenibacillus* genus. The Greengenes 13.5 database is used for demonstration purposes but the other `MgDb` annotation packages can also be used; RDP 11.5 - *ribosomaldatabaseproject11.5MgDb* or SILVA 128.1 - *silva128.1MgDb*.

5.6.2 Required Packages

In addition to *metagenomeFeatures* and *greengenes13.5MgDb* the *DECIPHER*, *tidyverse*, and *ggpubr* packages are also used in the following analysis. Our analysis uses the `DECIPHER` package to extract the amplicon regions, perform multiple sequence alignment, and generate a pairwise sequence distance matrix [**Wright2016-mo**]. The `tidyverse` and `ggpubr` packages will be used to reformat the taxonomic and distance matrix data and generate summary figures [106, 47].

```
library(tidyverse); packageVersion("tidyverse")
```

```
## [1] '1.2.1'
```

```
library(ggpubr); packageVersion("ggpubr")
```

```
## [1] '0.1.6'
```

```
library(DECIPHER); packageVersion("DECIPHER")
```

```
## [1] '2.8.1'
```

```
library(metagenomeFeatures); packageVersion("metagenomeFeatures")
```

```
## [1] '2.0.0'
```

```
library(greengenes13.5MgDb); packageVersion("greengenes13.5MgDb")
```

```
## [1] '2.0.0'
```

5.6.3 *Paenibacillus* Sequence and Taxonomy Data

We first subset the Greeengenes 13.5 database using the `mgDb_select` function. Then summarize the taxonomy data using functions from `tidyverse` package, specifically `dplyr`, `stringr` and `forcats` functions for manipulating `data.frames`, `strings`, and `factor` vectors respectively.

```
paeni_16S <- metagenomeFeatures::mgDb_select(gg13.5MgDb,
                          type = c("taxa","seq"),
                          keys = "Paenibacillus",
                          keytype = "Genus")
```

```
## Per genus count data
taxa_df <- paeni_16S$taxa %>%
    ## cleaning up species names
    mutate(Species = if_else(Species == "s__",
                             "Unassigned", Species),
           Species = str_replace(Species, "s__","")) %>%
    group_by(Species) %>%
    summarise(Count = n()) %>%
    ungroup() %>%
    mutate(Species = fct_reorder(Species, Count))

## Count info for text
total_otus <- sum(taxa_df$Count)
unassigned_idx <- taxa_df$Species == "Unassigned"
no_species_assignment <- taxa_df$Count[unassigned_idx]
```

For the Greengenes 13.5 database, there are a total of 2912 sequences classified as 15 species in the Genus *Paenibacillus*. The number of sequences assigned to specific *Paenibacillus* species range from 199 *Paenibacillus amylolyticus* to 2 *Paenibacillus illinoisensis* (Fig. 5.1). Sequences only classified to the genus level, "Unassigned", is the most abundant group, 2308.

5.6.4 Taxonomic resolution

Next, we evaluate the 16S rRNA amplicon sequencing taxonomic resolution for *Paenibacillus* species by comparing within and between species amplicon pairwise distance for the V12 and V4 regions. To differentiate between species the pairwise distances for within-species amplicon sequences must be less than the between species distances. Additionally ,the difference in amplicon sequence pairwise dis-

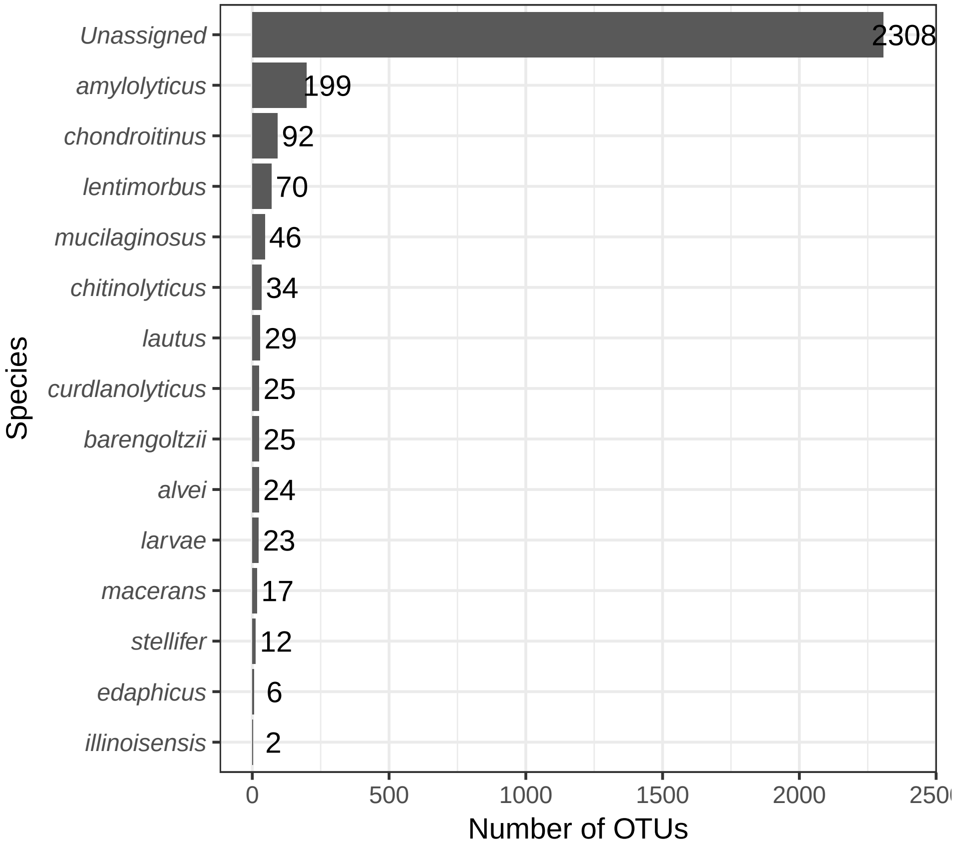


Figure 5.1: Number of sequences assigned to species in the genus *Paenibacillus*.

tances between and within species must be greater than the sequencing error rate to detect the difference. For our taxonomic resolution analysis, we used pattern matching to extract the V12 and V4 regions of the 16S rRNA sequences. We then generate a pairwise distance matrix for the two regions and compare the within and between species pairwise distances.

For our *in-silico* PCR we will use the following PCR primers:

Region	Direction	Primer
V12	Forward	27F - AGAGTTTGATCATGGCTCAG
	Reverse	336R - CACTGCTGCSYCCCGTAGGAGTCT
V4	Forward	515F - GTGCCAGCMGCCGCGGTAA
	Reverse	806R - GGACTACHVGGGTWTCTAAT

5.6.4.1 V12

Extracting the V12 region from the database sequences, only sequences with containing both forward and reverse primers are included in the analysis.

```
forward_primer <- "AGAGTTTGATCATGGCTCAG"
## reverse complementing reverse primer
reverse_primer <- DNAString("CACTGCTGCSYCCCGTAGGAGTCT") %>%
    reverseComplement() %>%
    as.character()

## Finding sequeces with forward primer
forward_match <- Biostrings::vmatchPattern(forward_primer,
                                           subject = paeni_16S$seq,
                                           max.mismatch = 2) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")

## Finding sequences with reverse primer
reverse_match <- Biostrings::vmatchPattern(reverse_primer,
                                           subject = paeni_16S$seq,
                                           max.mismatch = 2,
                                           fixed = FALSE) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")

## sequences with both forward and reverse primers
seqs_to_use_ids <- intersect(forward_match$seq_id,
                             reverse_match$seq_id)
seqs_to_use <- names(paeni_16S$seq) %in% seqs_to_use_ids

## Trimming sequences with both primers
paeni_V12 <- TrimDNA(paeni_16S$seq[seqs_to_use],
                     leftPatterns = forward_primer,
                     rightPatterns = reverse_primer,
                     type = "both")
```

```
## Finding left pattern: 100% internal, 0% flanking
##
## Finding right pattern: 100% internal, 0% flanking
##
## Time difference of 0.06 secs
```

```
## Excluding seqs with lenght 0
paeni_V12_seqs <- paeni_V12[[2]][width(paeni_V12[[2]]) != 0]
```

Generating a multiple sequence alignment using the `AlignSeqs` function in the `DECIPHER` package.

```
v12_align <- AlignSeqs(paeni_V12[[2]], verbose = FALSE)
```

The resulting alignment can be viewed using the `BrowseSeqs` function in the `DECIPHER` package.

```
BrowseSeqs(v12_align)
```

Generating pairwise distance matrix using the `DistanceMatrix` function in the `DECIPHER` package for taxonomic resolution analysis and converting distance matrix to a data frame for analysis.

```
v12_dist <- DistanceMatrix(v12_align,
                           correction = "none",
                           verbose = FALSE,
                           includeTerminalGaps = FALSE)

v12_dist_df <- v12_dist %>%
    as.data.frame() %>%
    rownames_to_column(var = "Keys") %>%
    gather("Keys2","distance", -Keys) %>%
    mutate(Keys = as.numeric(Keys),
           Keys2 = as.numeric(Keys2)) %>%
    filter(Keys < Keys2) %>%
    mutate(Keys = as.character(Keys),
           Keys2 = as.character(Keys2))

tax_df <- dplyr::select(paeni_16S$taxa, "Keys", "Species")
v12_dist_anno_df <- v12_dist_df %>%
    left_join(tax_df) %>%
    left_join(tax_df,by = c("Keys2" = "Keys")) %>%
    dplyr::rename(Keys_Species = Species.x,
                  Keys2_Species = Species.y) %>%
        mutate(group_comp = if_else(Keys_Species == Keys2_Species,
                                    "within","between")) %>%
    filter(Keys_Species != "s__", Keys2_Species != "s__")
```

5.6.4.2 V4

For the V4 region, we will use the same approach, extract amplicon region, filter extracted sequences based on amplicon length, generate pairwise distance matrix using a multiple sequence alignment, and then evaluate pairwise distances.

```
## Finding sequeces with forward primer
forward_match <- Biostrings::vmatchPattern("GTGCCAGCMGCCGCGGTAA",
                                           subject = paeni_16S$seq,
                                           fixed = FALSE) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")

## Finding sequences with reverse primer
reverse_match <- Biostrings::vmatchPattern("ATTAGAWACCCBDGTAGTCC",
                                           subject = paeni_16S$seq,
                                           fixed = FALSE) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")
```

```
## sequences with both forward and reverse primers
seqs_to_use_ids <- intersect(forward_match$seq_id,
                             reverse_match$seq_id)
seqs_to_use <- names(paeni_16S$seq) %in% seqs_to_use_ids

## Extract amplicon region
paeni_V4 <- TrimDNA(paeni_16S$seq[seqs_to_use],
                    leftPatterns = "GTGCCAGCMGCCGCGGTAA",
                    rightPatterns = "ATTAGAWACCCBDGTAGTCC",
                    type = "both")
```

```
## Finding left pattern: 100% internal, 0% flanking
##
## Finding right pattern: 100% internal, 0% flanking
##
## Time difference of 0.81 secs
```

```
## Excluding seqs with lenght 0
paeni_V4_seqs <- paeni_V4[[2]][width(paeni_V4[[2]]) != 0]

### Calculate distance matrix from multiple sequence alignment
v4_align <- AlignSeqs(paeni_V4_seqs, verbose = FALSE)
v4_dist <- DistanceMatrix(v4_align,
                          correction = "none",
                          verbose = FALSE,
                          includeTerminalGaps = FALSE)

## Creating a data frame for exploratory analysis
v4_dist_df <- v4_dist %>%
    as.data.frame() %>%
    rownames_to_column(var = "Keys") %>%
    gather("Keys2","distance", -Keys) %>%
    mutate(Keys = as.numeric(Keys), Keys2 = as.numeric(Keys2)) %>%
    filter(Keys < Keys2) %>%
    mutate(Keys = as.character(Keys), Keys2 = as.character(Keys2))

tax_df <- dplyr::select(paeni_16S$taxa, "Keys", "Species")
v4_dist_anno_df <- v4_dist_df %>%
    left_join(tax_df) %>%
    left_join(tax_df,by = c("Keys2" = "Keys")) %>%
    dplyr::rename(Keys_Species = Species.x,
                  Keys2_Species = Species.y) %>%
    mutate(group_comp = if_else(Keys_Species == Keys2_Species,
                                "within","between")) %>%
    filter(Keys_Species != "s__", Keys2_Species != "s__")

## Excluding outlier sequence "329842"
```

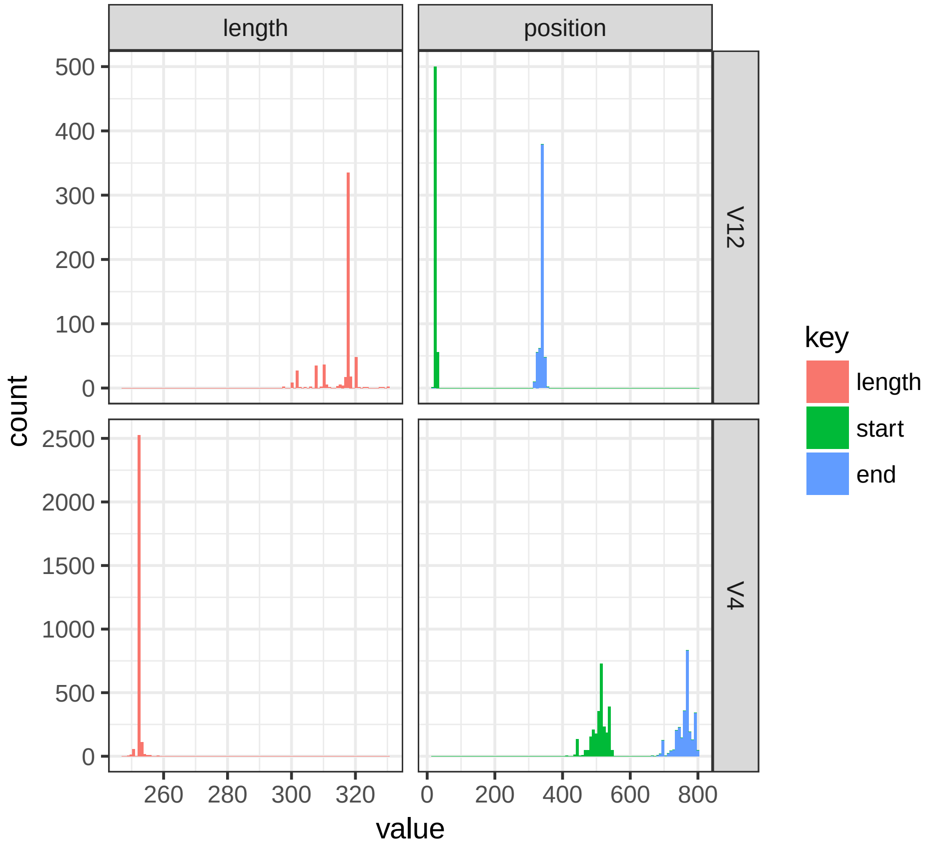


Figure 5.2: Primer trimmed sequence, amplicon, length and start and end positions relative to full length sequences for the V12 and V4 regions.

```
## - mean pairwise distance to all other
## sequences is 0.2
 v4_dist_anno_filt <- filter(v4_dist_anno_df,
                             Keys != "329842",
                             Keys2 != "329842")
```

5.6.4.3 Amplicon Sequence Lengths

The trimmed sequence length varies for forward and reverse primers resulting in varying amplicon sizes for both the V12 and V4 amplicons (Fig. 5.2).

Genus Level Comparison   Pairwise distance is significantly different for within and between species comparisons indicating that the V12 and V4 regions are potentially

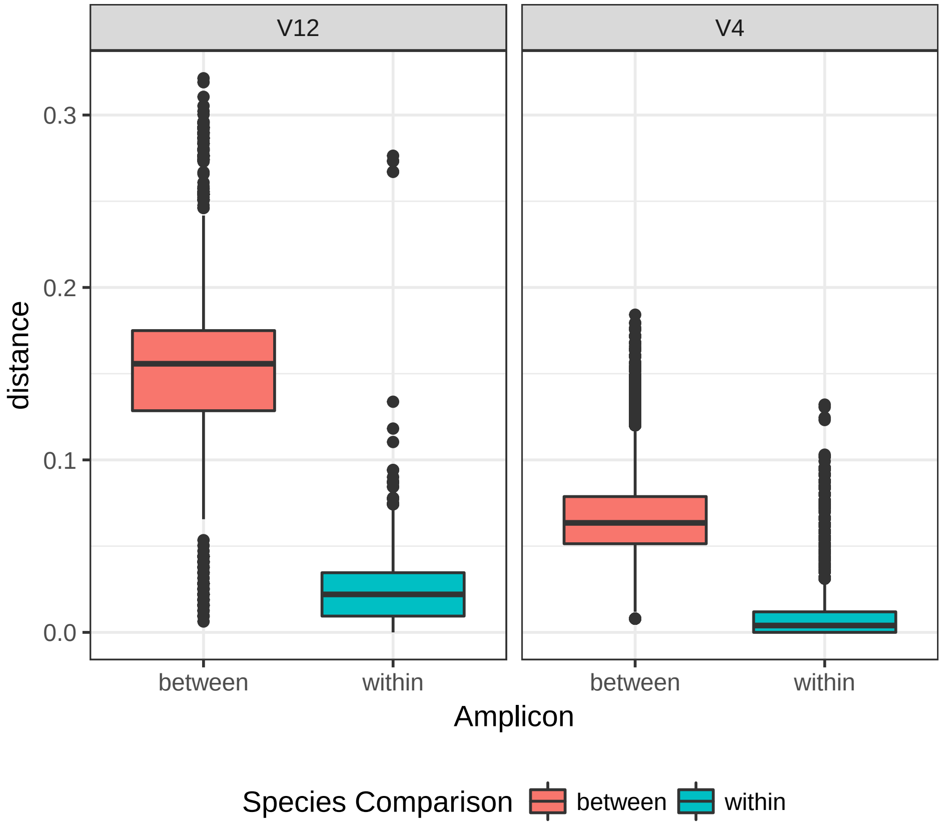


Figure 5.3: Distribution of within and between species pairwise distances for the V4 16S rRNA region. Sequences not classified to the species level were excluded from the analysis.

suitable forclassifying members of the *Paenibacillus* genus to the species level (Fig. 5.3). Overall the V12 region had greater pairwise distances than V4 for both within and between species. It is important also to consider that the majority of sequences in the database were only classified to the genus level. Species-level information for these sequences might yield results that are inconsistent with our analysis. Additionally, our analysis does not identify the pairwise sequence distance required to classify a sequence as a novel *Paenibacillus* species.

Species level comparison While the overall pairwise distance is greater between species than within species for the *Paenibacillus* genus, it is important to understand how the within and between species pairwise distances compare for individual species. The heatmap below shows pairwise distance information for within and between different *Paenibacillus* species for the V12 and V4 regions (Fig. 5.4). Whether the sequences are assigned to more than one OTU depends on the pairwise sequence distance metric and linkage method employed by the clustering algorithm. In general though for species levels classification the maximum within species distance should be less than the minimum between species distance. For example as the maximum within species pairwise distance for *P. lentimorbus* is 0.13 and the minimum between species pairwise distance for *P. lentimorbus* and *P. alvei* is 0.08 (Fig. 5.4A), correctly assigning a V12 amplicon sequences to one of these two species is not possible.

5.6.5 Conclusion

Here we demonstrate how the *metagenomeFeatures* package in conjunction with one of the associated 16S rRNA database packages, *greengenes13.5MgDb*, and other R packages, can be used to evaluate whether species-level taxonomic classification is possible for a specific amplicon region. The approach used here can easily be extended to use different 16S rRNA databases (starting with a different `MgDb`class object), taxonomic groups (changing filtering parameters), or amplicon regions (changing primer sequences).

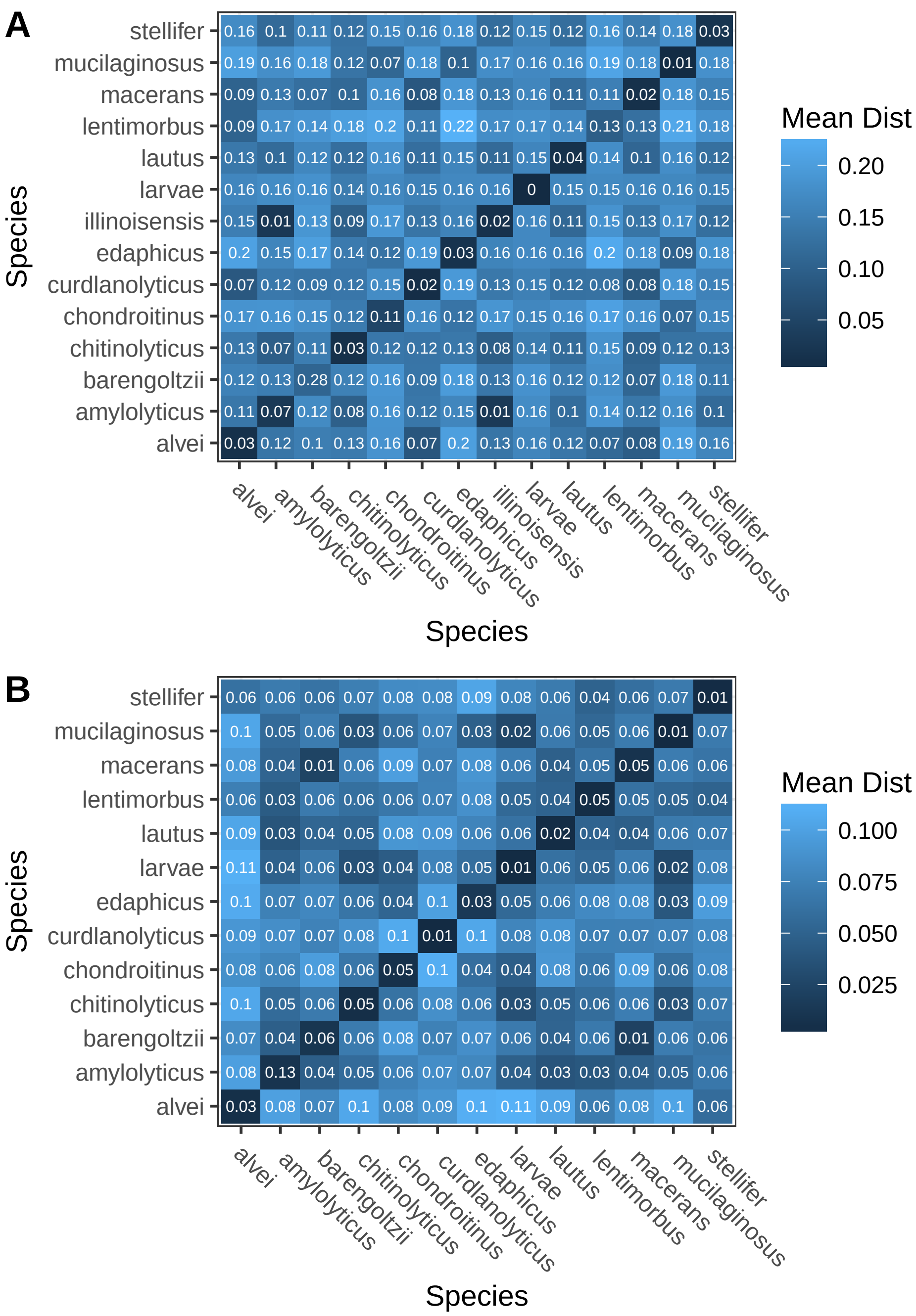


Figure 5.4: Pairwise distances between *Paenibacillus* species (A) V12 and (B) V4 amplicon regions. Fill color indicates the mean pairwise sequence distance within and between species. The text indicates the maximum pairwise distance for within-species comparisons, values along the diagonal, and maximum pairwise distance for between species comparisons. Different number of species are included in the V12 and V4 plots as there are no full-length *P. chondroitinus* sequences with the V12 primer in the database.

CHAPTER 6

Conclusions

For this dissertation, I developed a framework for assessing the 16S rRNA marker-gene survey measurement process. The framework utilizes novel statistical methods in conjunction with an assessment dataset specifically developed for this dissertation. I created mixtures of human gut microbiome samples and sequenced them in multiple laboratories and runs. Based on this experimental design I defined multiple measurement assessment metrics. The statistical methods assess 16S rRNA marker-gene survey relative abundance, differential abundance, and beta diversity using information from the unmixed samples and mixture design. The mixture dataset was a two-sample titration series of vaccine trial DNA extracts. Additionally, I developed the R Bioconductor package, *metagenomeFeatures* for working with 16S rRNA reference databases and marker-gene survey feature data.

The *metagenomeFeatures* package provides data structures and functions for working with 16S rRNA gene sequence reference databases and marker-gene survey feature data. The data structure provided by the `MgDb-class` in conjunction with the shared sequence identifier system developed by RNACentral facilitates comparisons between 16S rRNA databases. The `mgFeatures-class` provides the groundwork for the development of a common data structure for working with metagenomic and marker-gene sequence data in R which will increase interoperability between R packages developed for working with metagenomic sequence data. Additionally, while the data structures were developed for 16S rRNA gene sequence data they can be used for any marker-gene sequence data without modification and can be extended to work with shotgun metagenomic sequence data and databases.

Based on lessons learned from this dissertation, new mixture datasets can be developed for further microbiome measurement assessment. Additional 16S rRNA sequencing mixture datasets would serve as a complementary resource for the community. Using samples with either better characterized prokaryotic DNA proportions, or minimal non-prokaryotic DNA, would reduce the expected value uncertainty observed in this assessment. Also, using samples with larger differences in microbial composition as titration endpoints to generate the mixtures would provide a more extensive set of features for assessment. As the mixtures were only processed using a single laboratory protocol (16S PCR through sequencing), mixture samples can be used as part of an interlaboratory study to further characterize the measurement process repeatability and reproducibility. Finally, mixtures can be used to assess other microbiome measurement processes such as shotgun metagenomics, metatranscriptomics, and even non-nucleic acid measurements such as metaproteomics and metametabolomics.

The work presented here has shown how a mixture dataset can be used to assess the marker-gene survey measurement process. Using the assessment framework, I evaluated 16S rRNA marker-gene survey bioinformatic pipeline and normalization performance. Bioinformatic pipelines combine multiple algorithms converting raw sequence data into count tables which are subsequently used to test biological hypotheses. Algorithm choice and parameters can significantly impact pipeline results. The pipelines compared in this dissertation were optimized using mock communities and benchmarked against other methods based on similarity in beta-diversity results [10]. The assessment framework and dataset provide complementary methods for use in optimizing existing and benchmarking new pipelines and normalization methods. The mixture dataset can be processed with any bioinformatic pipeline that converts raw 16S rRNA sequencing data to a count table. The relative and differential abundance, as well as beta-diversity assessment, can be performed on the count table and the results compared to those obtained with the pipelines evaluated in this dissertation. Future work includes the development of an R Bioconductor package for employing our assessment framework, *metagenomeAssessment*.

Bibliography

[1] Amnon Amir et al. “Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns”. en. In: *mSystems* 2.2 (Mar. 2017).

[2] Marti J Anderson et al. “Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist”. en. In: *Ecol. Lett.* 14.1 (Jan. 2011), pp. 19–28.

[3] Erik Aronesty. “ea-utils: Command-line tools for processing biological sequencing data”. In: *Expression Analysis, Durham, NC* (2011).

[4] Shawn C Baker et al. “The external RNA controls consortium: a progress report”. In: *Nature methods* 2.10 (2005), pp. 731–734.

[5] Monika Balvoit and Daniel H Huson. “SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare?” en. In: *BMC Genomics* 18.Suppl 2 (Mar. 2017), p. 114.

[6] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: `10.18637/jss.v067.i01`.

[7] Robert G Beiko. “Microbial malaise: how can we classify the microbiome?” en. In: *Trends Microbiol.* 23.11 (Nov. 2015), pp. 671–679.

[8] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[9] BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML. *The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. JCGM.* Tech. rep. 200. 2012.

[10] Nicholas A Bokulich et al. “mockrobiota: a public resource for microbiome bioinformatics benchmarking”. In: *mSystems* 1.5 (2016), e00062–16.

[11] Daniel Borcard, Pierre Legendre, and Pierre Drapeau. “Partialling out the spatial component of ecological variation”. In: *Ecology* 73.3 (1992), pp. 1045–1055.

[12] Hans W. Borchers. *pracma: Practical Numerical Math Functions.* R package version 2.1.4. 2018. URL: `https://CRAN.R-project.org/package=pracma`.

[13] J Roger Bray and J T Curtis. "An Ordination of the Upland Forest Communities of Southern Wisconsin". In: *Ecol. Monogr.* 27.4 (Feb. 1957), pp. 325–349.

[14] J Paul Brooks et al. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies". In: *BMC Microbiol.* 15.1 (2015), pp. 1–14.

[15] J Paul Brooks et al. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies". In: *BMC microbiology* 15.1 (2015), p. 66.

[16] R. Brent Calder. *savR: Parse and analyze Illumina SAV files.* R package version 1.14.0. 2015. URL: `https://github.com/bcalder/savR`.

[17] Benjamin Callahan. *Silva taxonomic training data formatted for DADA2 (Silva version 128).* These files are derived from the Silva database, and their use is subject to the Silva dual- licensing model for academia and commercial users: https://www.arb-silva.de/silva-license- information/. July 2017. DOI: `10.5281/zenodo.824551`. URL: `https://doi.org/10.5281/zenodo.824551`.

[18] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis". en. In: *ISME J.* 11.12 (Dec. 2017), pp. 2639–2643.

[19] Benjamin J Callahan et al. "DADA2: High-resolution sample inference from Illumina amplicon data". In: *Nature Methods* 13 (2016), pp. 581–583. DOI: `10.1038/nmeth.3869`.

[20] BJ Callahan et al. "Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; referees: 3 approved]". In: *F1000Research* 5.1492 (2016). DOI: `10.12688/f1000research.8986.2`.

[21] J. Gregory Caporaso et al. "QIIME allows analysis of high-throughput community sequencing data". In: *Nature Methods* 7 (Apr. 2010). Correspondence, p. 335. URL: `http://dx.doi.org/10.1038/nmeth.f.303`.

[22] Tsute Chen et al. "The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information". en. In: *Database* 2010 (July 2010), baq013.

[23] Ilseung Cho and Martin J Blaser. "The human microbiome: at the interface of health and disease". In: *Nature Reviews Genetics* 13.4 (2012), p. 260.

[24] Jinlyung Choi et al. "Strategies to improve reference databases for soil microbiomes". en. In: *ISME J.* 11.4 (Apr. 2017), pp. 829–834.

[25] Adam G Clooney et al. "Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis". In: *PLoS One* 11.2 (2016), e0148028.

[26] James R Cole et al. "Ribosomal Database Project: data and tools for high throughput rRNA analysis". en. In: *Nucleic Acids Res.* 42.Database issue (Jan. 2014), pp. D633–42.

[27] Rosalinda D'Amore et al. "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling". In: *BMC Genomics* 17 (2016), pp. 1–40. ISSN: 1471-2164. DOI: 10.1186/s12864-015-2194-9. URL: http://dx.doi.org/10.1186/s12864-015-2194-9.

[28] Todd Z DeSantis et al. "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB". In: *Applied and environmental microbiology* 72.7 (2006), pp. 5069–5072.

[29] Robert C Edgar. "Search and clustering orders of magnitude faster than BLAST". In: *Bioinformatics* 26.19 (2010), pp. 2460–2461.

[30] Robert C Edgar. "UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing". In: *BioRxiv* (2016), p. 081257.

[31] Robert C Edgar. "Updating the 97% identity threshold for 16S ribosomal RNA OTUs". In: *bioRxiv* (2017), p. 192211.

[32] Robert C Edgar et al. "UCHIME improves sensitivity and speed of chimera detection". In: *Bioinformatics* 27.16 (2011), pp. 2194–2200.

[33] SLR Ellison and A Williams. "Eurachem/CITAC guide: Quantifying Uncertainty in Analytical Measurement, ISBN 978-0-948926-30-3". In: *Available from www.eurachem.org* (2013).

[34] Julia K Goodrich et al. "Conducting a microbiome study". In: *Cell* 158.2 (2014), pp. 250–262.

[35] Nicholas J Gotelli and Robert K Colwell. "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness". In: *Ecol. Lett.* 4.4 (July 2001), pp. 379–391.

[36] Elliot Nicholas Grady et al. "Current knowledge and perspectives of Paenibacillus: a review". In: *Microbial cell factories* 15.1 (2016), p. 203.

[37] Micah Hamady, Catherine Lozupone, and Rob Knight. "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data". en. In: *ISME J.* 4.1 (Jan. 2010), pp. 17–27.

[38] Clayton Harro et al. "Refinement of a human challenge model for evaluation of enterotoxigenic Escherichia coli vaccines". In: *Clinical and Vaccine Immunology* 18.10 (2011), pp. 1719–1727.

[39] Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature Methods* 12.2 (2015), pp. 115–121. URL: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

[40] Jennifer B Hughes and Jessica J Hellmann. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity". In: *Methods in Enzymology.* Vol. 397. Academic Press, Jan. 2005, pp. 292–308.

[41] Human Microbiome Project Consortium. "Structure, function and diversity of the healthy human microbiome". en. In: *Nature* 486.7402 (June 2012), pp. 207–214.

[42] Susan M Huse et al. "Ironing out the wrinkles in the rare biosphere through improved OTU clustering." In: *Environmental microbiology* 12.7 (July 2010), pp. 1889–98. ISSN: 1462-2920. DOI: 10.1111/j.1462-2920.2010.02193.x. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2909393&tool=pmcentrez&rendertype=abstract.

[43] Rafael A Irizarry et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". In: *Biostatistics* 4.2 (2003), pp. 249–264.

[44] Paul Jaccard. "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1". In: *New Phytol.* 11.2 (Feb. 1912), pp. 37–50.

[45] J Michael Janda and Sharon L Abbott. "16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls". In: *Journal of clinical microbiology* 45.9 (2007), pp. 2761–2764.

[46] Jens Kallmeyer et al. "Global distribution of microbial abundance and biomass in subseafloor sediment". In: *Proceedings of the National Academy of Sciences* 109.40 (2012), pp. 16213–16216.

[47] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.1.6. 2017. URL: https://CRAN.R-project.org/package=ggpubr.

[48] Anna Klindworth et al. "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies". In: *Nucleic acids research* (2012), gks808.

[49] Heidi H Kong et al. "Performing Skin Microbiome Research: A Method to the Madness". en. In: *J. Invest. Dermatol.* 137.3 (Mar. 2017), pp. 561–568.

[50] Evguenia Kopylova et al. "Open-Source Sequence Clustering Methods Improve the State Of the Art". In: *mSystems* 1.1 (2014), pp. 1–16. ISSN: 2379-5077. DOI: 10.1128/mSystems.00003-15.Editor.

[51] James J Kozich et al. "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform". In: *Applied and environmental microbiology* 79.17 (2013), pp. 5112–5120.

[52] Justin Kuczynski et al. "Experimental and analytical tools for studying the human microbiome". In: *Nature Reviews Genetics* 13.1 (2012), p. 47.

[53] Jenna M Lang et al. "A microbial survey of the International Space Station (ISS)". en. In: *PeerJ* 5 (Dec. 2017), e4029.

[54] Ruth E Ley et al. "Microbial ecology: human gut microbes associated with obesity". In: *Nature* 444.7122 (2006), p. 1022.

[55] B Magnusson and U Ornemark. "Eurachem/CITAC guide: The Fitness for Purpose of Analytical Methods–A Laboratory Guide to Method Validation and Related Topics (2nd ed.). ISBN 978-91-87461-59-0". In: *Available from www.eurachem.org* (2014).

[56] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. en. In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12.

[57] Marcel Martinez-Porchas et al. “How conserved are the conserved 16S-rRNA regions?” en. In: *PeerJ* 5 (Feb. 2017), e3036.

[58] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. en. In: *Nucleic Acids Res.* 40.10 (May 2012), pp. 4288–4297.

[59] Daniel McDonald et al. “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea”. en. In: *ISME J.* 6.3 (Mar. 2012), pp. 610–618.

[60] Paul J McMurdie and Susan Holmes. “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data”. en. In: *PLoS One* 8.4 (Apr. 2013), e61217.

[61] Paul J McMurdie and Susan Holmes. “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data”. en. In: *PLoS One* 8.4 (Apr. 2013), e61217.

[62] Paul J McMurdie and Susan Holmes. “Waste not, want not: why rarefying microbiome data is inadmissible.” In: *PLoS Comput. Biol.* 10.4 (Apr. 2014), e1003531.

[63] Martin Morgan et al. “ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data”. In: *Bioinformatics* 25 (2009), pp. 2607–2608. DOI: `10.1093/bioinformatics/btp450`. URL: `http://dx.doi.org10.1093/bioinformatics/btp450`.

[64] Kirill Müller et al. *’SQLite’ Interface for R [R package RSQLite version 2.0]*. Comprehensive R Archive Network (CRAN), 2017.

[65] Nam-phuong Nguyen et al. “TIPP: taxonomic identification and phylogenetic profiling”. In: *Bioinformatics* 30.24 (2014), pp. 3548–3555.

[66] Jari Oksanen et al. *vegan: Community Ecology Package*. R package version 2.4-6. 2018. URL: `https://CRAN.R-project.org/package=vegan`.

[67] Jie Ouyang et al. “Paenibacillus thiaminolyticus: a new cause of human infection, inducing bacteremia in a patient on hemodialysis”. In: *Annals of Clinical & Laboratory Science* 38.4 (2008), pp. 393–400.

[68] H Pagès, M Lawrence, and P Aboyoun. “S4Vectors: S4 implementation of vectors and lists”. In: *R package version 0. 13* 15 (2017).

[69] H Pagès et al. “Biostrings: String objects representing biological sequences, and matching algorithms”. In: *R package version* 2 (2008), p. 160.

[70] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. “APE: Analyses of Phylogenetics and Evolution in R language”. en. In: *Bioinformatics* 20.2 (Jan. 2004), pp. 289–290.

[71] Jerod Parsons et al. “Using mixtures of biological samples as process controls for RNA-sequencing experiments”. In: *BMC genomics* 16.1 (2015), p. 708.

[72] Joseph N Paulson et al. “Differential abundance analysis for microbial marker-gene surveys”. In: *Nature methods* 10.12 (2013), pp. 1200–1202.

[73] Joseph N Paulson et al. “Differential abundance analysis for microbial marker-gene surveys”. en. In: *Nat. Methods* 10.12 (Dec. 2013), pp. 1200–1202.

[74] Anna Y Pei et al. “Diversity of 16S rRNA genes within individual prokaryotic genomes”. In: *Applied and environmental microbiology* 76.12 (2010), pp. 3886–3897.

[75] James B Pettengill and Hugh Rand. “Segal’s Law, 16S rRNA gene sequencing, and the perils of foodborne pathogen detection within the American Gut Project”. en. In: *PeerJ* 5 (June 2017), e3480.

[76] P Scott Pine, Barry A Rosenzweig, and Karol L Thompson. “An adaptable method using human mixed tissue ratiometric controls for benchmarking performance on gene expression microarrays in clinical laboratories”. In: *BMC biotechnology* 11.1 (2011), p. 38.

[77] Mihai Pop et al. “Individual-specific changes in the human gut microbiota after challenge with enterotoxigenic Escherichia coli and subsequent ciprofloxacin treatment”. In: *BMC genomics* 17.1 (2016), p. 1.

[78] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. “FastTree 2–approximately maximum-likelihood trees for large alignments”. en. In: *PLoS One* 5.3 (Mar. 2010), e9490.

[79] Christian Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic acids research* 41.D1 (2012), pp. D590–D596.

[80] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: `https://www.R-project.org/`.

[81] Krishna Rao and Vincent B Young. “Fecal microbiota transplantation for the management of Clostridium difficile infection”. In: *Infectious Disease Clinics* 29.1 (2015), pp. 109–122.

[82] Jai Ram Rideout et al. “Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences”. en. In: *PeerJ* 2 (Aug. 2014), e545.

[83] Alice Risely et al. “Gut microbiota of a long-distance migrant demonstrates resistance against environmental microbe incursions”. In: *Molecular ecology* (2017).

[84] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. en. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140.

[85] Robin Rebecca Rohwer et al. "TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution". In: *bioRxiv* (Jan. 2017).

[86] Melanie Schirmer et al. "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform". In: *Nucleic Acids Research* 43.6 (2015). ISSN: 13624962. DOI: `10.1093/nar/gku1341`.

[87] Schliep et al. "Intertwining phylogenetic trees and networks". In: *Methods in Ecology and Evolution* 8.10 (2017), pp. 1212–1220. ISSN: 2041-210X. DOI: `10.1111/2041-210X.12760`. URL: `http://dx.doi.org/10.1111/2041-210X.12760`.

[88] Patrick D Schloss and Jo Handelsman. "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness". en. In: *Appl. Environ. Microbiol.* 71.3 (Mar. 2005), pp. 1501–1506.

[89] Patrick D Schloss et al. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities". In: *Applied and environmental microbiology* 75.23 (2009), pp. 7537–7541.

[90] Ron Sender, Shai Fuchs, and Ron Milo. "Revised estimates for the number of human and bacteria cells in the body". In: *PLoS biology* 14.8 (2016), e1002533.

[91] Luke Sheneman, Jason Evans, and James A Foster. "Clearcut: a fast implementation of relaxed neighbor joining". en. In: *Bioinformatics* 22.22 (Nov. 2006), pp. 2823–2824.

[92] Rashmi Sinha et al. "Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium". en. In: *Nat. Biotechnol.* 35.11 (Nov. 2017), pp. 1077–1086.

[93] Welliton Souza and Benilton Carvalho. *Rqc: Quality Control Tool for High-Throughput Sequencing Data.* R package version 1.10.2. 2017. URL: `https://github.com/labbcb/Rqc`.

[94] Francesco Strati et al. "New evidences on the altered gut microbiota in autism spectrum disorders". In: *Microbiome* 5.1 (2017), p. 24.

[95] Marc A Sze and Patrick D Schloss. "Looking for a signal in the noise: revisiting obesity and the microbiome". In: *MBio* 7.4 (2016), e01018–16.

[96] The RNAcentral Consortium. "RNAcentral: a comprehensive database of non-coding RNA sequences". en. In: *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D128–D134.

[97] Karol L Thompson et al. "Use of a mixed tissue RNA design for performance assessments on multiple microarray formats". In: *Nucleic acids research* 33.22 (2005), e187–e187.

[98] Luke R Thompson et al. "A communal catalogue reveals Earth's multiscale microbial diversity". en. In: *Nature* 551.7681 (Nov. 2017), pp. 457–463.

[99] Peter J Turnbaugh et al. “The human microbiome project”. In: *Nature* 449.7164 (2007), p. 804.

[100] William Walters et al. “Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys”. en. In: *mSystems* 1.1 (Jan. 2016).

[101] Qiong Wang et al. “Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy”. In: *Applied and environmental microbiology* 73.16 (2007), pp. 5261–5267.

[102] Sophie Weiss et al. “Normalization and microbial differential abundance strategies depend upon data characteristics”. en. In: *Microbiome* 5.1 (Mar. 2017), p. 27.

[103] Sarah L Westcott and Patrick D Schloss. “OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units”. In: *mSphere* 2.2 (2017).

[104] William B Whitman, David C Coleman, and William J Wiebe. “Prokaryotes: the unseen majority”. In: *Proceedings of the National Academy of Sciences* 95.12 (1998), pp. 6578–6583.

[105] Hadley Wickham. *A 'dplyr' Back End for Databases [R package dbplyr version 1.1.0]*. Comprehensive R Archive Network (CRAN), 2017.

[106] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 2017. URL: `https://CRAN.R-project.org/package=tidyverse`.

[107] Hadley Wickham et al. *A Grammar of Data Manipulation [R package dplyr version 0.7.4]*. Comprehensive R Archive Network (CRAN), 2017.

[108] Erik S. Wright. “Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R”. In: *The R Journal* 8.1 (2016), pp. 352–359.

[109] Erik S Wright et al. “Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates”. In: *Environmental microbiology* 16.5 (2014), pp. 1354–1365.

[110] Bo Yang, Yong Wang, and Pei-Yuan Qian. “Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis”. In: *BMC bioinformatics* 17.1 (2016), p. 1.

[111] Vincent B Young. “The role of the microbiome in human health and disease: an introduction for clinicians”. In: *BMJ* 356 (2017), j831.