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

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.

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 (Goodrich et al. 2014). 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 (Goodrich et al. 2014; Kong et al. 2017).

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 (Schloss and Handelsman 2005) and closed reference clustering of reads against a reference database (Edgar 2010). 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 (Rideout et al. 2014). 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 (B. J. Callahan et al. 2016; Amir et al. 2017). 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 (Gotelli and Colwell 2001). Although there are concerns about its statistical validity (McMurdie and Holmes 2014), rarefaction is currently the only normalization method for unweighted,

presence-absence based, beta-diversity metrics (Weiss et al. 2017). 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 (Paulson et al. 2013). 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 (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012), 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 (Hamady, Lozupone, and Knight 2010). 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 (Bray and Curtis 1957; Jaccard 1912). 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 (Anderson et al. 2011).

Previous studies have evaluated different bioinformatics pipelines (Sinha et al. 2017) and normalization methods (McMurdie and Holmes 2014; Weiss et al. 2017) 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.

3. Methods

Our assessment framework utilizes a dataset of DNA mixtures from five vaccine trial participants described in Section **??**. 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. 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 (Calder 2015) and ShortRead Bioconductor R packages (Morgan et al. 2009).

3.0.1. *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. Pre-processing and feature detection methods vary by pipeline. The Mothur pipeline uses the OptiClust algorithm for *de novo* clustering (Westcott and Schloss 2017). Pre-processing includes merging and quality filtering paired-end reads followed by aligning sequences to the SILVA reference alignment (Schloss et al. 2009). Taxonomic classification was performed using the RDP Bayesian classifier (Wang et al. 2007) implemented in Mothur. The phylogenetic tree was constructed in Mothur using the clearcut algorithm (Sheneman, Evans, and Foster 2006). 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 (Cole et al. 2014).

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/) (Martin 2011). 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 (B. J. Callahan et al. 2016) and formatted SILVA database version 128 trainset provided by the DADA2 developers (Callahan 2017). Taxonomic classification was performed using the DADA2 implementation of the RDP Bayesian classifier (Wang et al. 2007). The phylogenetic tree was generated following methods in (BJ Callahan et al. 2016) using the DECIPHER R package for multiple sequence alignment (Wright 2016) and the phangorn R package for tree construction (Schliep et al. 2017).

The QIIME pipelines all used the same merged paired-end, quality filtered set of sequences (Caporaso et al. 2010). UCLUST alogrithm (version v1.2.22q) was used for clustering and taxonomic assignment against the Greengenes database version 13.8 97% similarity OTUs (Edgar 2010; McDonald et al. 2012). Phylogenetic trees were constructed using FastTree, and a multiple sequence alignment generated using pyNAST and the Greengenes reference alignment (Caporaso et al. 2010; Price, Dehal, and Arkin 2010). 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) (Amir et al. 2017). Phylogenetic tree construction methods used for the other QIIME pipelines were also used for the Deblur pipeline.

3.0.2. *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 (Gotelli and Colwell 2001; Hughes and Hellmann 2005). 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 (Thompson et al. 2017) and other comparison studies (Weiss et al. 2017; McMurdie and Holmes 2014). 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 (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012), and those that are commonly used to normalize 16S rRNA marker-gene survey, such as cumulative sum scaling (CSS) (Paulson et al. 2013) 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 (McMurdie and Holmes 2013). Weighted and unweighted UniFrac phylogenetic beta-diversity metrics were calculated using the phyloseq implementation of FastUniFrac (McMurdie and Holmes 2013; Hamady, Lozupone, and Knight 2010). For feature-level beta-diversity assessment, the Bray-Curtis weighted, and Jaccard unweighted metrics were used (Bray and Curtis 1957; Jaccard 1912).

3.0.3. *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 (Bates et al. 2015). 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.

3.0.3.1. 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.

3.0.3.2. 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. 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 (Borchers 2018). 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.

3.0.3.3. 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 (Borcard, Legendre, and Drapeau 1992) to quantify technical and biological factor's contribution to the total observed variation. Variation partition was calculated using the Vegan R package (Oksanen et al. 2018). Distance-based redundancy analysis (dbRDA) was used to identify significant sources of variation (Oksanen et al. 2018).

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. 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.

4.1. **Dataset Characteristics.** The four replicate sequencing runs were of variable sequence quality and depth (Fig. 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 reverse reads (Fig. 2A). Read base quality was lower for the reverse read than the forward reads for all four sequencing runs (Fig. 2B). Sequence data from the two NIST runs had higher quality scores than the data from JHU runs, except for JHU2 forward reads (Fig. 2B). Greater variability in sample feature total abundance was observed on the first run at each laboratory (Fig. 2C).

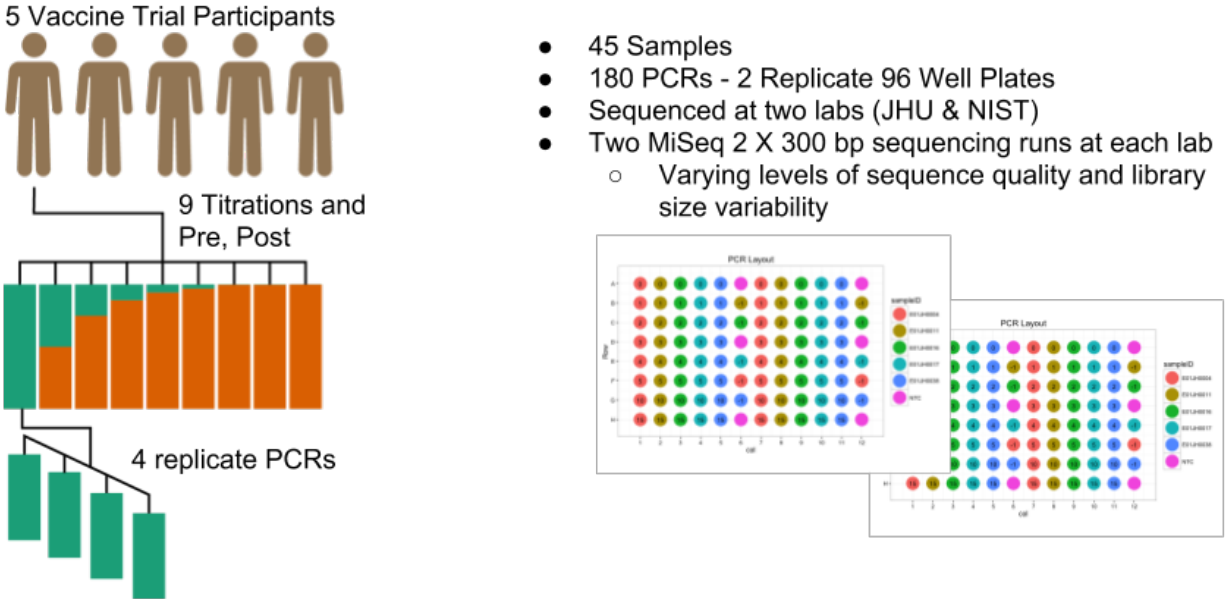

FIGURE 1. Two-sample titration dataset experimental design. The dataset contained independent two-sample titration series from 5 vaccine trial participants (subjects), resulting in 45 samples. PCRs were run on two 96 well plates with each plate half containing one for each sample and three no template 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 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)

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.

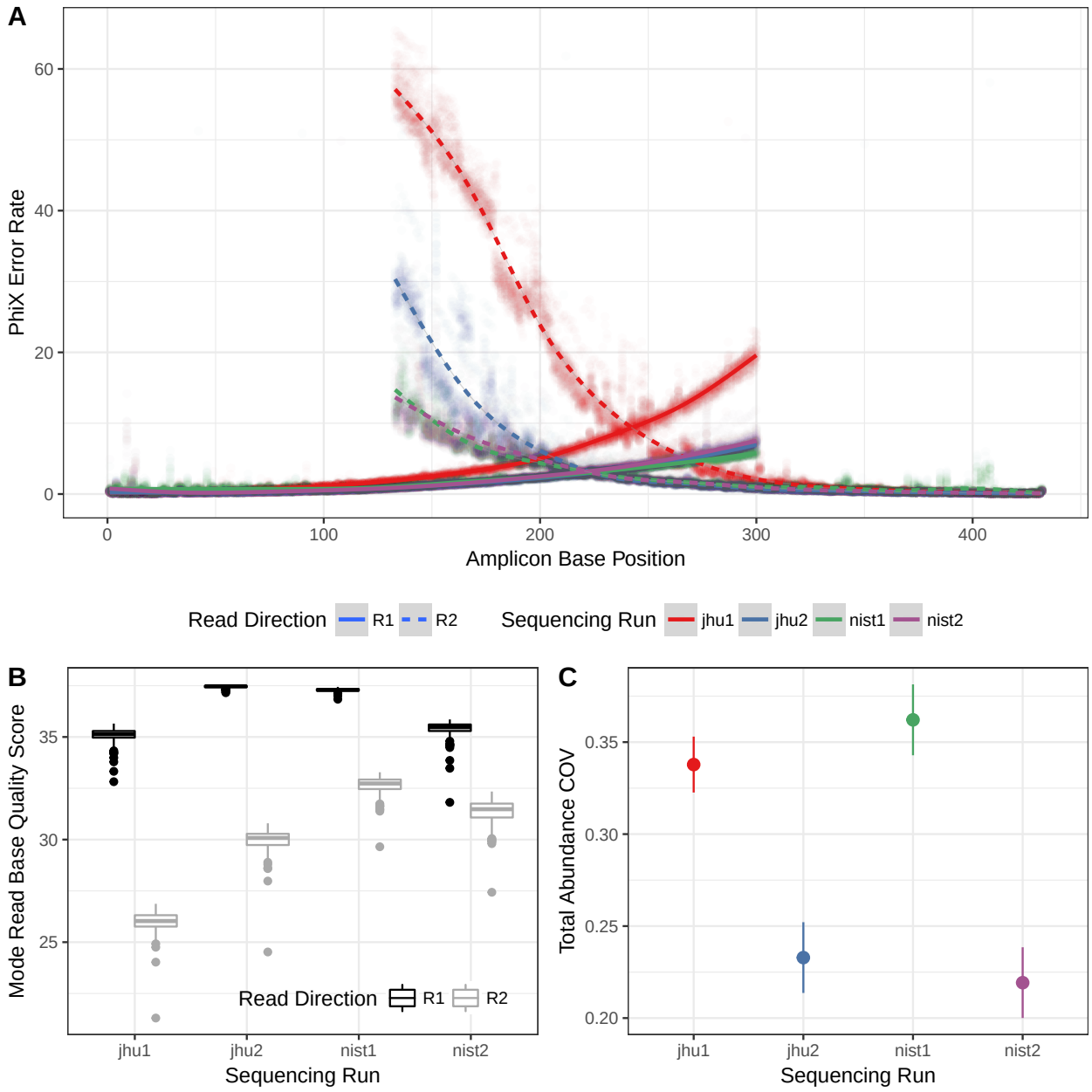

FIGURE 2. Sequencing quality and sample total abundance variation for the four sequencing runs used in this study. The same set of 192 PCRs were sequenced in all four runs. Independent sequencing libraries were generated at the two sequencing laboratories (JHU and NIST). (A) PhiX error rate relative 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.

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.

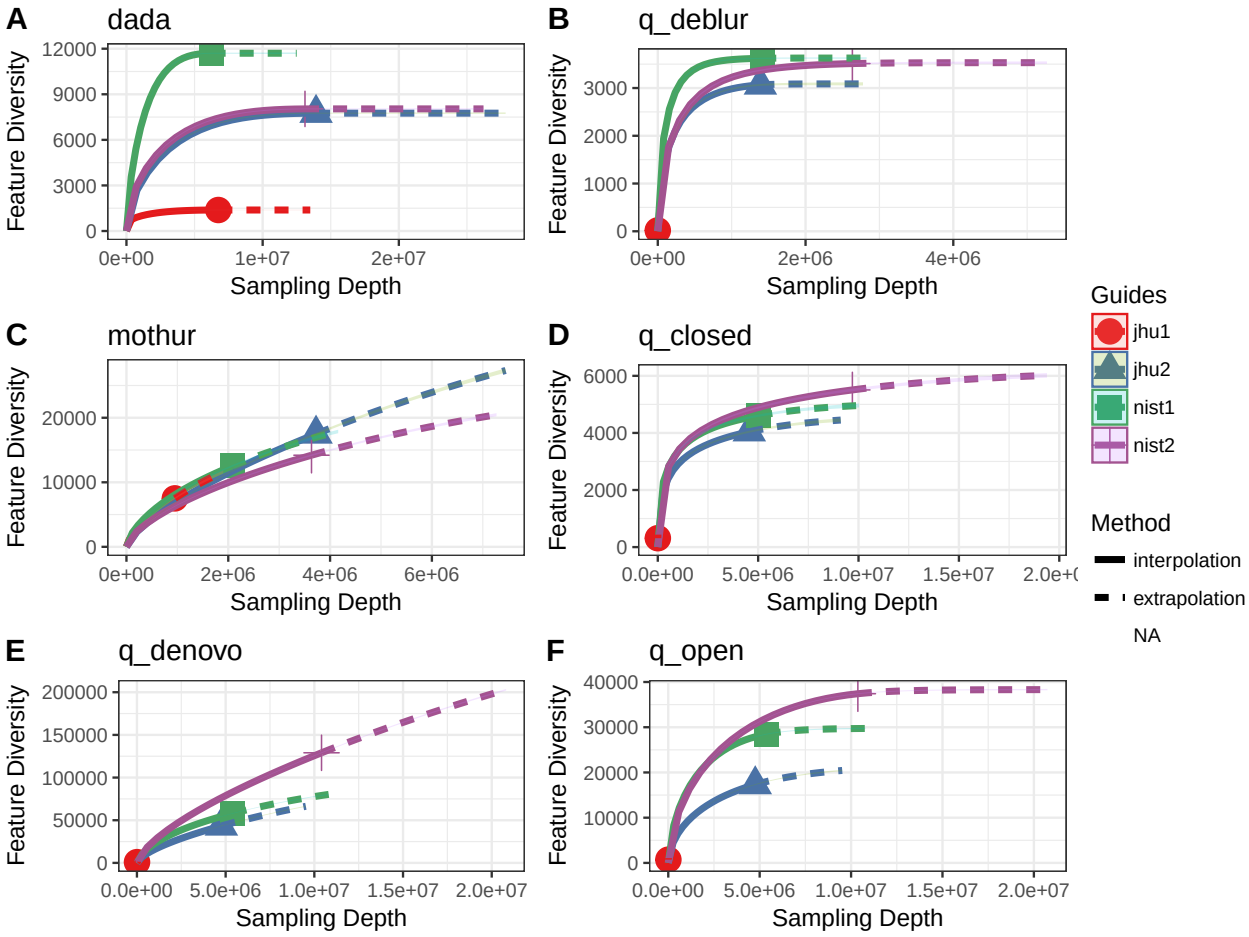

FIGURE 3. Rarefaction curves for the four sequencing runs (line color) by pipeline (A-F). Rarefaction curves were calculated using the feature counts summed across all samples by sequencing run. Rarefaction curves indicate how thoroughly a population is sampled. Curves show the relationship between the 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.

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 1).

We generated rarefaction curves to assess feature diversity at multiple sampling depths for across the four sequencing runs (Fig. 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. 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. 3A). Rarefaction curves for *de novo*, open-reference, and closed-reference methods did not reach an asymptote (Fig. 3). The QIIME *de novo* pipeline had the greatest slope, suggesting the highest rate of artifacts (Fig. 3E). This is most likely due to

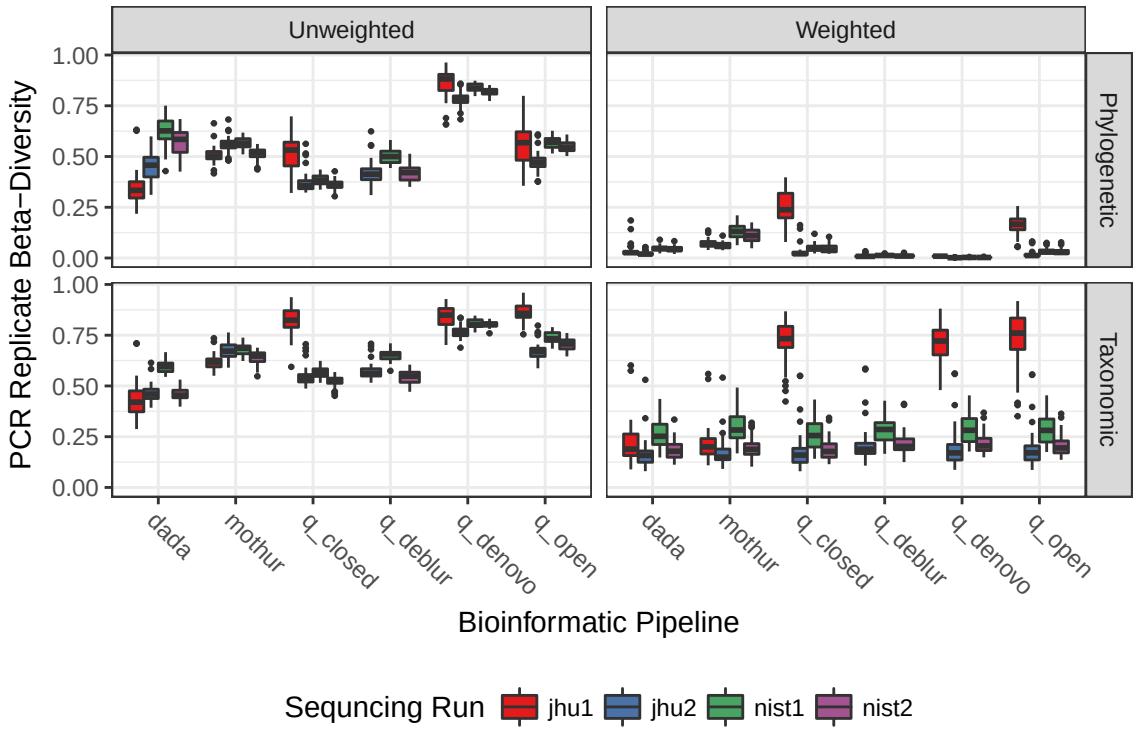

FIGURE 4. Distribution of mean pairwise PCR replicate beta-diversity by sequencing run and pipeline for un-normalized count data.

the fact that the QIIME *de novo* pipeline does not filter out singletons (Table 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. 3D-F).

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. 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 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. 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 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 PCR replicate beta-diversity, we observed that most normalization methods reduced beta-diversity between PCR replicates (Fig. 5A). For a number of pipelines, TMM and RLE normalization methods significantly lowered weighted PCR replicate beta-diversity (Fig. 5A). For unweighted metrics (Fig. 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.

4.3. **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

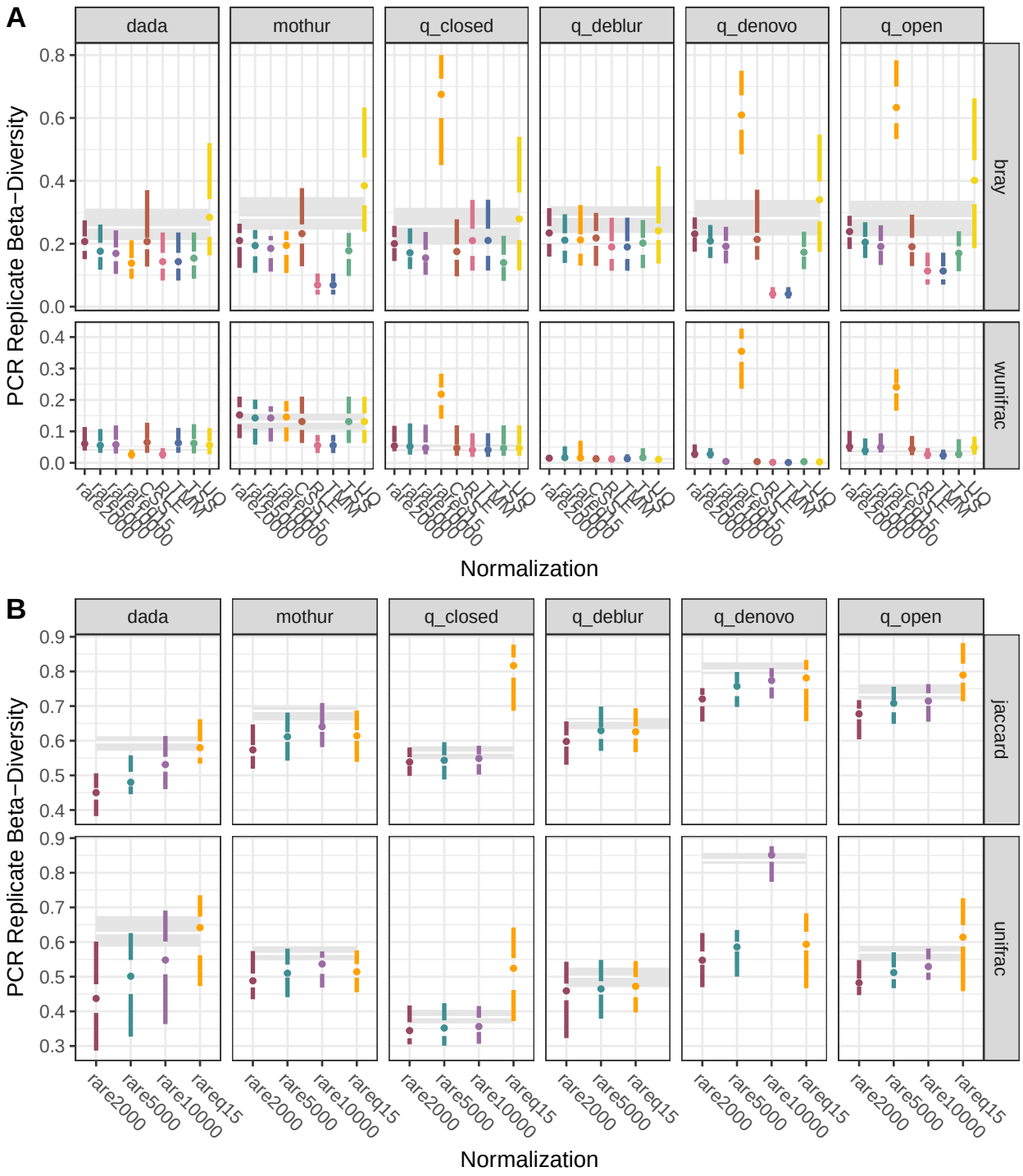

FIGURE 5. Impact of normalization method on mean weighted (A) and unweighted (B) PCR replicates beta-diversity, for the sequencing run with higher quality and total abundance variability, NIST1. Data are presented as minimal-ink boxplots, where points indicate median value, the gap between point and 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.

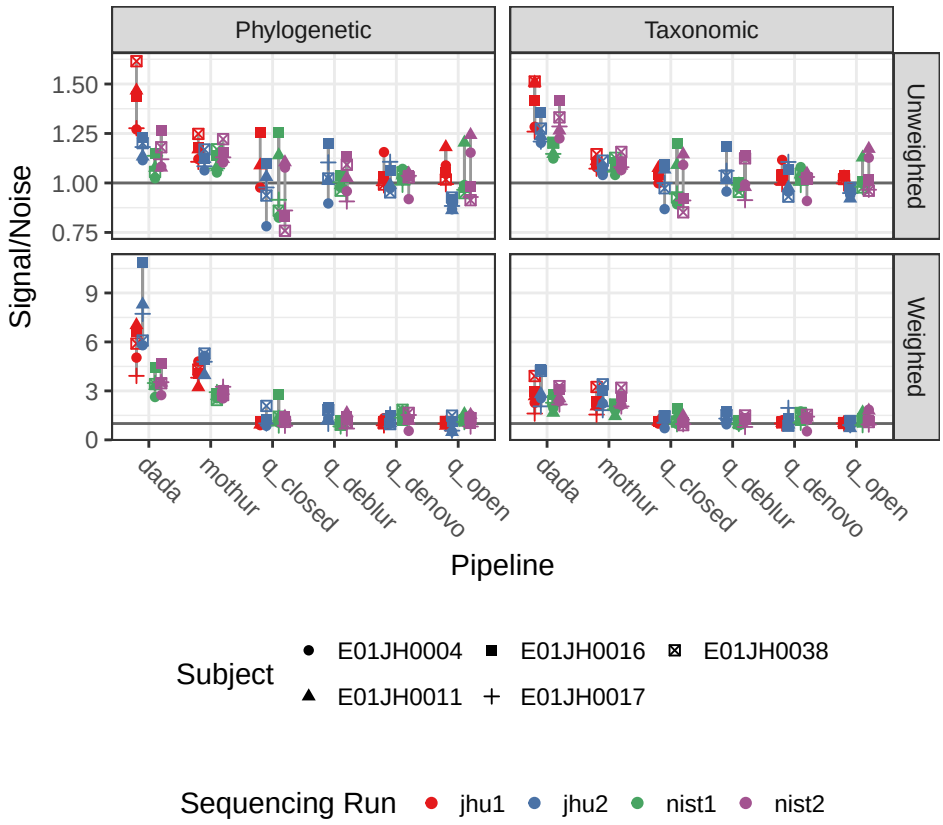

FIGURE 6. The weighted average signal to noise varied by pipeline, run, and diversity metric. Points indicate the signal to noise for each individual with grey lines representing the range of values for a pipeline and sequencing run. Dark grey horizontal lines indicate a signal-to-noise ratio of 1.

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. 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 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. 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. 7B), though the increase was only significant at lower subsampling depths for DADA2 and Mothur pipelines.

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. 8). The magnitude of this difference, however, was greater for weighted than unweighted beta-diversity metrics

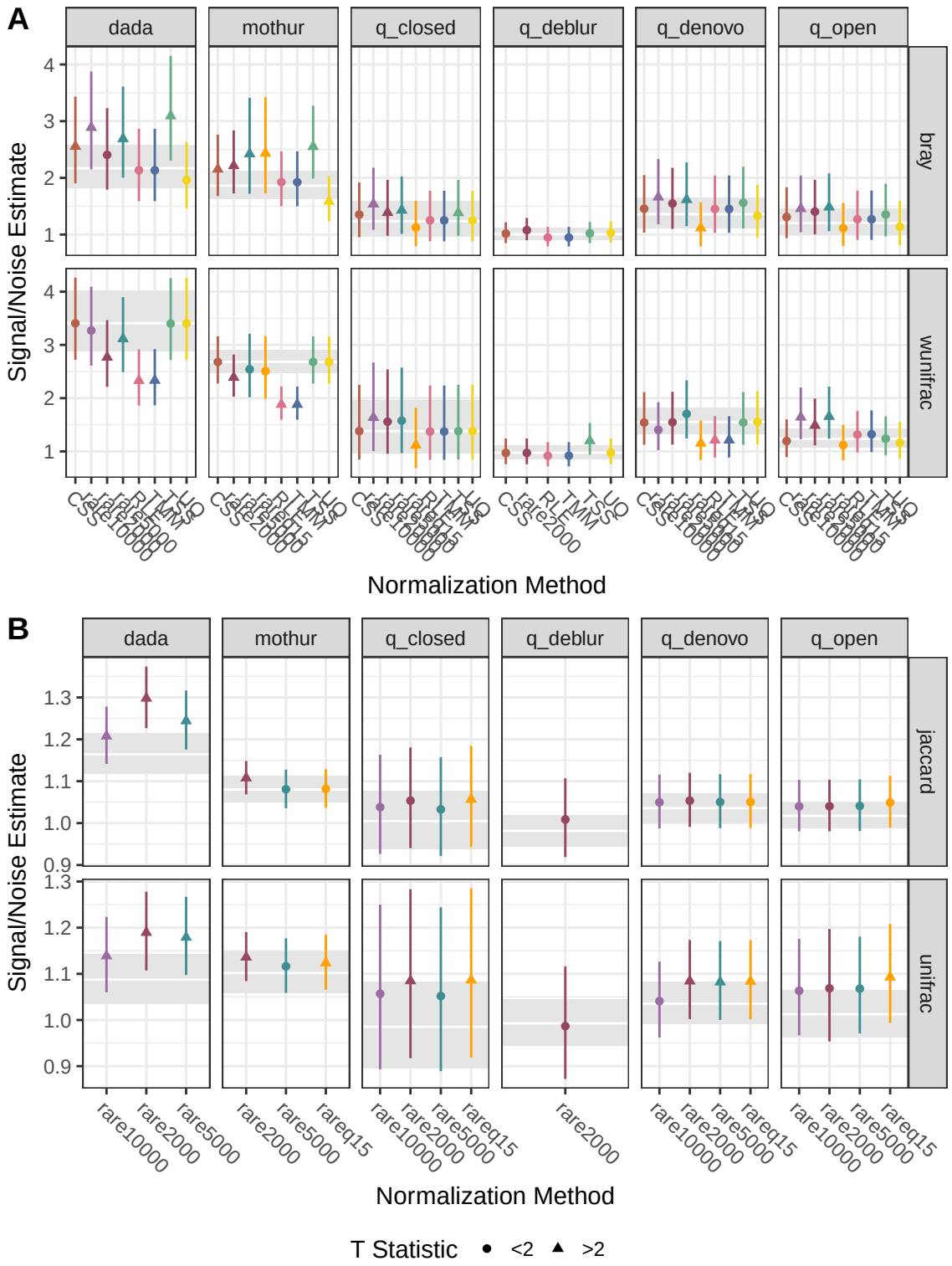

FIGURE 7. Weighted average signal to noise ratio estimate and 95 CI for raw and normalized count data for (A) weighted and (B) unweighted beta-diversity metrics. Estimates were calculated using a mixed effects linear model using subject as random effect. The horizontal solid line is the unnormalized 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.

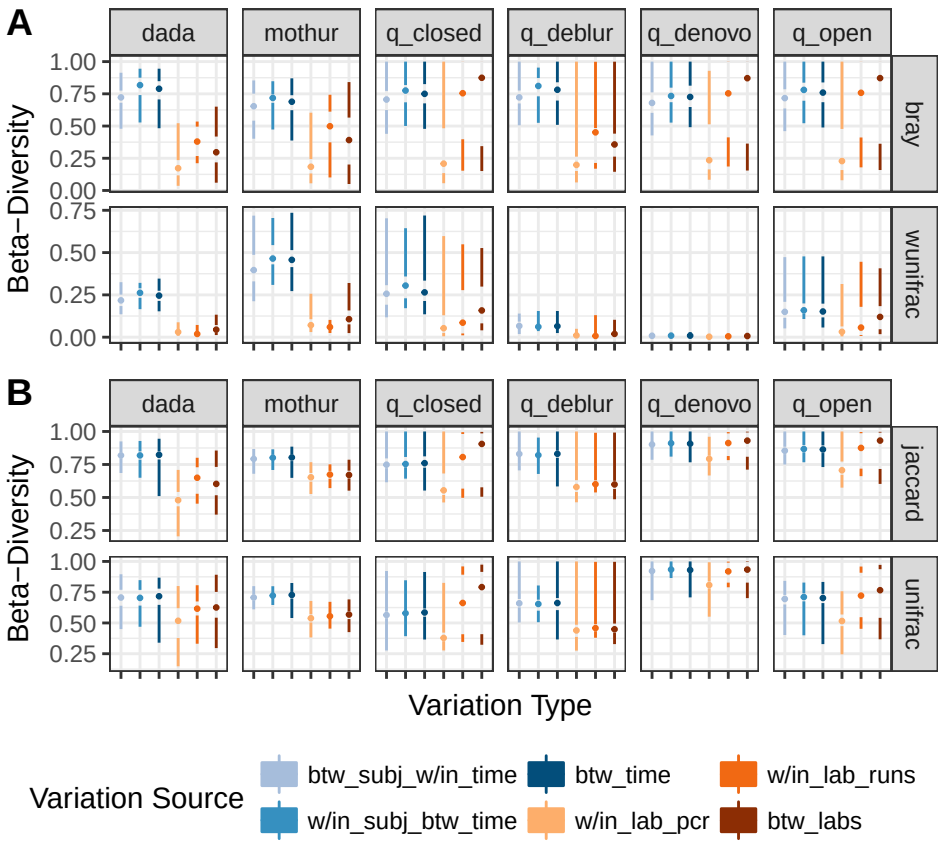

FIGURE 8. Biological vs. Technical Variation, distribution in (A) weighted and (B) unweighted beta-diversity between technical replicates and biological treatments (subject and timepoint).

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 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. 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. 9A) and unweighted (Fig. 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. 9A). Principal coordiante analysis plots for the unmixed pre-exposure samples are consistent

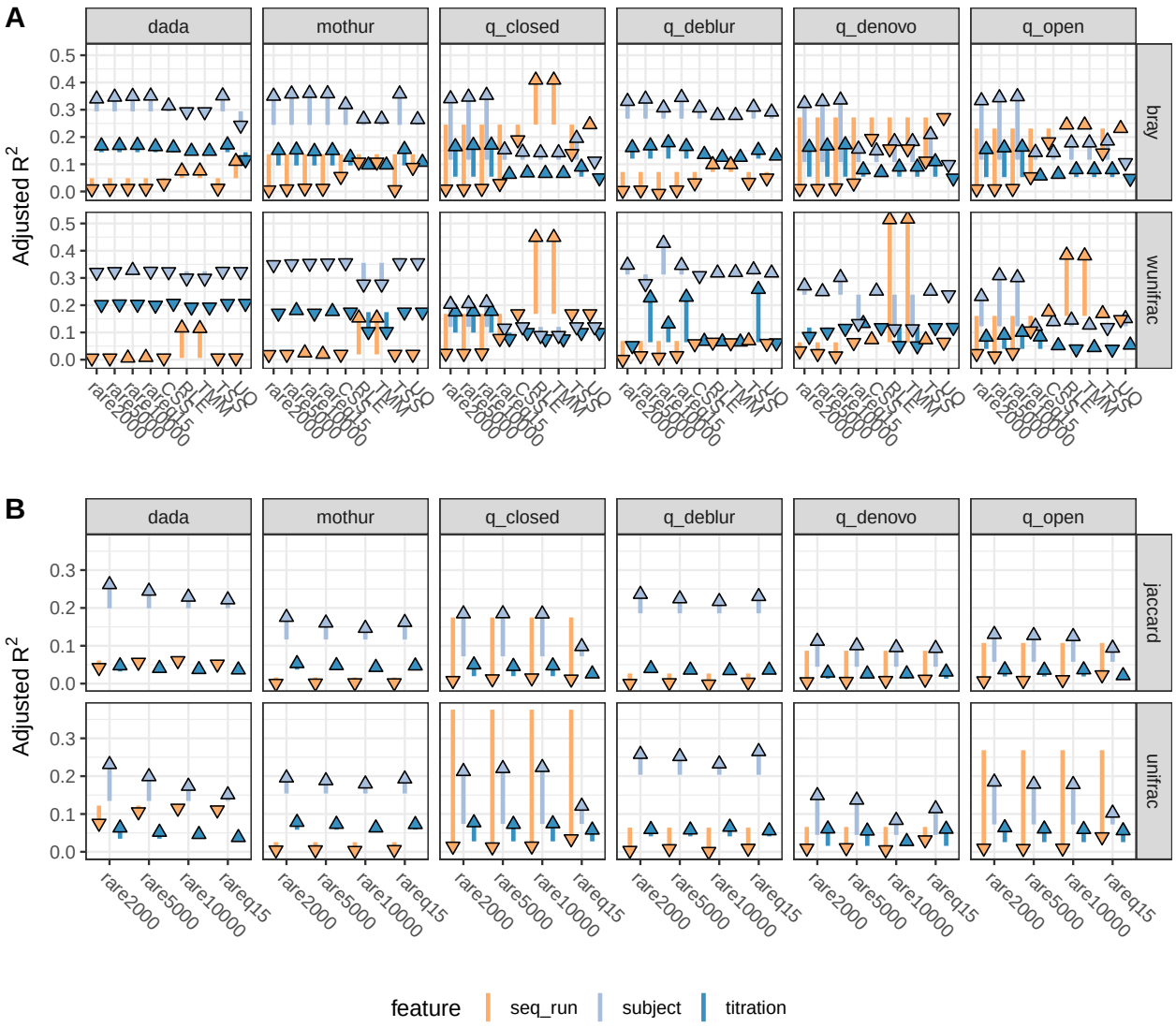

FIGURE 9. Impact of different normalization methods on biological and technical sources of variatoin for different pipelines and beta-diversity metrics. y-axis is the adjusted R^2, indicating the proportion of variance explained by each biological (subject and titration) and technical (seq run) variable. 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.

with variation partitioning results (Fig. 10). For Mothur and DADA2 the technical replicates group more tightly when TSS is used to normalize count data compared to when TMM.

5. Discussion

Sequence error rate and variation in library size are just two sequencing characteristics that can negatively bias beta-diversity analyses (McMurdie and Holmes 2014). Ideally, bioinformatic pipelines can help differentiate true biological sequences from artifacts generated by sequencing errors (B. J. Callahan et al. 2016) and normalization methods, such as rarefaction and total sum scaling, can adjust for differences in library size (Paulson et al. 2013). 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

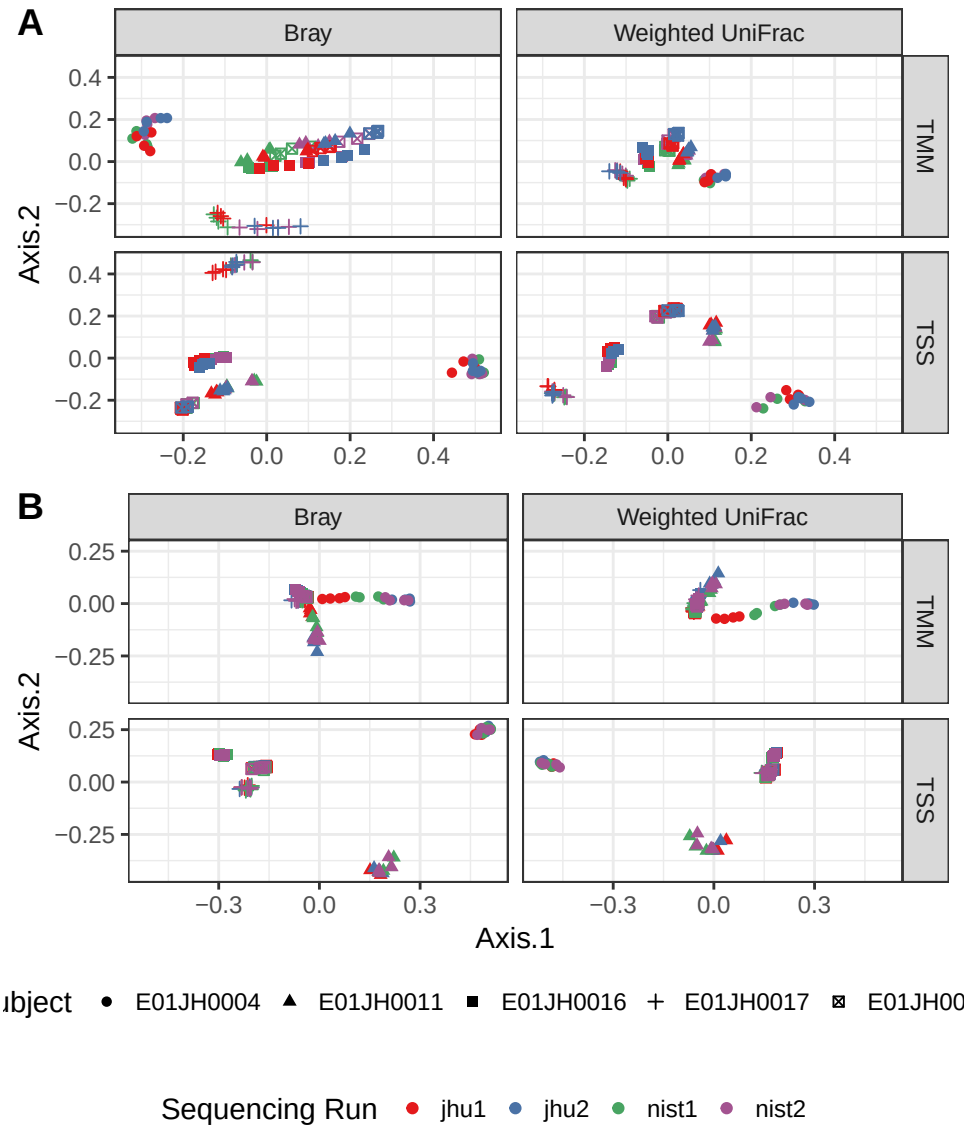

FIGURE 10. Principal coordinate analysis for TMM and TSS normalized (A) DADA2 and (B) Mothur unmixed PRE samples for Bray-Curtis and Weighted UniFrac distance metrics.

TABLE 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	-	-	-

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 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 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. 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. 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 (McMurdie and Holmes 2014), they may not appropriate for beta-diversity analysis.

5.1. **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 (McMurdie and Holmes 2014). 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 (Bokulich et al. 2016). The novel assessment framework and dataset presented here provides complementary methods for use in optimizing existing and benchmarking new pipelines and normalization methods.

6. References

Amir, Amnon, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2 (2).

Anderson, Marti J, Thomas O Crist, Jonathan M Chase, Mark Vellend, Brian D Inouye, Amy L Freestone, Nathan J Sanders, et al. 2011. "Navigating the Multiple Meanings of β Diversity: A Roadmap for the Practicing Ecologist." *Ecol. Lett.* 14 (1): 19–28.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bokulich, Nicholas A, Jai Ram Rideout, William G Mercurio, Arron Shiffer, Benjamin Wolfe, Corinne F Maurice, Rachel J Dutton, Peter J Turnbaugh, Rob Knight, and J Gregory Caporaso. 2016. "Mockrobiota: A Public Resource for Microbiome Bioinformatics Benchmarking." *mSystems* 1 (5). Am Soc Microbiol: e00062–16.

Borcard, Daniel, Pierre Legendre, and Pierre Drapeau. 1992. "Partialling Out the Spatial Component of Ecological Variation." *Ecology* 73 (3). Wiley Online Library: 1045–55.

Borchers, Hans W. 2018. *Pracma: Practical Numerical Math Functions.* https://CRAN.R-project.org/package=pracma.

Bray, J Roger, and J T Curtis. 1957. "An Ordination of the Upland Forest Communities of Southern Wisconsin." *Ecol. Monogr.* 27 (4). Ecological Society of America: 325–49.

Calder, R. Brent. 2015. *SavR: Parse and Analyze Illumina Sav Files.* https://github.com/bcalder/savR.

Callahan, Benjamin. 2017. "Silva taxonomic training data formatted for DADA2 (Silva version 128)." https://doi.org/10.5281/zenodo.824551.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13: 581–83. https://doi.org/10.1038/nmeth.3869.

Callahan, BJ, K Sankaran, JA Fukuyama, PJ McMurdie, and SP Holmes. 2016. "Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses [Version 2; Referees: 3 Approved]." *F1000Research* 5 (1492). https://doi.org/10.12688/f1000research.8986.2.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (April). Nature Publishing Group SN -: 335. http://dx.doi.org/10.1038/nmeth.f.303.

Cole, James R, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje. 2014. "Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis." *Nucleic Acids Res.* 42 (Database issue): D633–42.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than Blast." *Bioinformatics* 26 (19). Oxford University Press: 2460–1.

Goodrich, Julia K, Sara C Di Rienzi, Angela C Poole, Omry Koren, William A Walters, J Gregory Caporaso, Rob Knight, and Ruth E Ley. 2014. "Conducting a Microbiome Study." *Cell* 158 (2). Elsevier: 250–62.

Gotelli, Nicholas J, and Robert K Colwell. 2001. "Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness." *Ecol. Lett.* 4 (4). Blackwell Science Ltd: 379–91.

Hamady, Micah, Catherine Lozupone, and Rob Knight. 2010. "Fast UniFrac: Facilitating High-Throughput Phylogenetic Analyses of Microbial Communities Including Analysis of Pyrosequencing and PhyloChip Data." *ISME J.* 4 (1): 17–27.

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

Jaccard, Paul. 1912. "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1." *New Phytol.* 11 (2). Blackwell Publishing Ltd: 37–50.

Kong, Heidi H, Björn Andersson, Thomas Clavel, John E Common, Scott A Jackson, Nathan D Olson, Julia A Segre, and Claudia Traidl-Hoffmann. 2017. "Performing Skin Microbiome Research: A Method to the Madness." *J. Invest. Dermatol.* 137 (3): 561–68.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Res.* 40 (10): 4288–97.

McDonald, Daniel, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. 2012. "An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea." *ISME J.* 6 (3): 610–18.

McMurdie, Paul J, and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PLoS One* 8 (4): e61217.

———. 2014. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." *PLoS Comput. Biol.* 10 (4): e1003531.

Morgan, Martin, Simon Anders, Michael Lawrence, Patrick Aboyoun, Hervé Pagès, and Robert Gentleman. 2009. "ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-Throughput Sequence Data." *Bioinformatics* 25: 2607–8. https://doi.org/10.1093/bioinformatics/btp450.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2018. *Vegan: Community Ecology Package*. https://CRAN.R-project.org/package=vegan.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12). Nature Research: 1200–1202.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS One* 5 (3): e9490.

Rideout, Jai Ram, Yan He, Jose A Navas-Molina, William A Walters, Luke K Ursell, Sean M Gibbons, John Chase, et al. 2014. "Subsampled Open-Reference Clustering Creates Consistent, Comprehensive OTU Definitions and Scales to Billions of Sequences." *PeerJ* 2 (August): e545.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Schliep, Klaus, Potts, Alastair J., Morrison, David A., Grimm, and Guido W. 2017. "Intertwining Phylogenetic Trees and Networks." *Methods in Ecology and Evolution* 8 (10): 1212–20. https://doi.org/10.1111/2041-210X.12760.

Schloss, Patrick D, and Jo Handelsman. 2005. "Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness." *Appl. Environ. Microbiol.* 71 (3): 1501–6.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23). Am Soc Microbiol: 7537–41.

Sheneman, Luke, Jason Evans, and James A Foster. 2006. "Clearcut: A Fast Implementation of Relaxed Neighbor Joining." *Bioinformatics* 22 (22): 2823–4.

Sinha, Rashmi, Galeb Abu-Ali, Emily Vogtmann, Anthony A Fodor, Boyu Ren, Amnon Amir, Emma Schwager, et al. 2017. "Assessment of Variation in Microbial Community Amplicon Sequencing by the Microbiome Quality Control (MBQC) Project Consortium." *Nat. Biotechnol.* 35 (11): 1077–86.

Thompson, Luke R, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681): 457–63.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16). Am Soc Microbiol: 5261–7.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, et al. 2017. "Normalization and Microbial Differential Abundance Strategies Depend Upon Data Characteristics." *Microbiome* 5 (1): 27.

Westcott, Sarah L, and Patrick D Schloss. 2017. "OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units." *mSphere* 2 (2).

Wright, Erik S. 2016. "Using Decipher V2.0 to Analyze Big Biological Sequence Data in R." *The R Journal* 8 (1): 352–59.