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

1. Abstract

Microbial communities, microbiomes, play an critical role in human and ecosystem health. 16S rRNA marker-gene sequencing is the most commonly used methods for characterizing microbiomes. Beta diversity metrics are commonly used to analyze microbial communities, previously developed for macro-ecology, characterize overall community similarity. The impact of sequence characteristics such as sequencing errors and differences in the number of reads generated per sample, library size, on beta-diversity analysis is not well understood. Bioinformatic pipelines and normalization methods are used to account for sequencing errors and library size differences. In the following study we assessed the impact of sequence characteristics on beta-diversity analysis, and how well different bioinformatic pipelines and normalization methods affect this impact. For this assessment we used a novel dataset consisting of stool samples from a vaccine trial participants, were samples collected before and after exposure to the pathogen were mixed following a two-sample titration. Multiple levels of replicates were included in the study, biological replicates (five vaccine trial participants) and technical replicates included PCR, sequencing library, and sequencing runs. The sequencing data were processed using six bioinformatic pipelines; DADA2 (sequence inference), Mothur (*de novo*), Deblur (QIIME 1 preprocessing), QIIME *de novo*, QIIME open-reference, and QIIME closed-reference. Normalization methods including multiple rarefying level, total sum scaling (TSS), cumulative sum scaling (CSS), upper quartile (UQ), trimmed mean of M values (TMM), and relative log expression (RLE). The assessment framework developed for this study consists of three components. (1) Beta-diversity repeatability for PCR replicates. (2) Difference in beta diversity between biological (e.g. different individuals) and technical factors (e.g. different sequencing runs). (3) Ability to differentiate groups of samples with varying levels of similarity, using titrations. The assessment results varied by pipeline, and normalization method. Mothur and DADA2 were less susceptible to sequencing errors. Ability of normalization methods to account for differences in sequencing depth varied by beta diversity metric and to less of an extent pipeline. Normalization methods TMM and RLE, developed for microarray and RNAseq data, are not appropriate for marker-gene survey beta-diversity analysis. For unweighted metrics, rarefying to 10,000 reads improved results whereas rarefying to the 85th percentile worsened results. While low error rates and consistent library size are ideal, we show that for beta-diversity analysis some bioinformatic pipelines and normalization methods are robust to lower quality sequence data.

2. Introduction

Microbial communities are frequently characterized via marker-gene surveys 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 improve our ability to resolve the taxonomy and diversity of microbiota, they are subject to biases that can significantly affect our interpretation of the resulting data. Bioinformatic pipelines and normalization methods are often used to reduce these biases, especially for beta diversity calculations comparing community structure between samples (Goodrich et al. 2014; Kong et al. 2017).

Bioinformatic pipelines reduce bias by remove sequencing artifacts from microbiome datasets. Sequence artifacts include single and multi- base pair variants, and chimeric sequences formed by the incorrect merging of two distinct biological molecules during PCR amplification. If not accounted for, these artifacts may incorrectly be attributed as novel diversity in a sample. Bioinformatic pipelines also perform clustering or sequence inference to group reads into biologically informative units. Standard clustering techniques include de-novo clustering based on pairwise similarities of sequences (Schloss and Handelsman 2005) and closed reference clustering of reads against a reference database (Edgar et al. 2011). Open reference clustering is a combination of the two, applying closed reference clustering first, 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 less abundant sequence is a sequencing artifact originating from the higher abundant sequence (B. J. Callahan

et al. 2016; Amir et al. 2017). The resulting features, OTUs (operational taxonomic units) for clustering methods and SVs (sequence variants) for sequence inference methods have different characteristics and vary in their ability to remove different types of sequence artifacts from the dataset, 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 macro ecology, where counts for a unit (sample) are randomly subsampled to a user defined constant level (Gotelli and Colwell 2001). While the statistical validity of rarifying is questionable (McMurdie and Holmes 2014), rarefaction is currently the only normalization method for unweighted, presence-absence based beta-diversity metrics (Weiss et al. 2017). Numeric normalization methods include total and cumulative sum scaling (TSS and CSS), where counts are divided by sample total abundance (TSS) or by the cumulative abundance for a defined percentile (Paulson et al. 2013). CSS is one of the few normalization methods developed with 16S rRNA marker-gene survey data in mind. 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 their suitability for beta diversity analysis is unclear.

Beta diversity is calculated using a variety of metrics that can be grouped based on whether they incorporate phylogenetic distance between features or not and whether they take into account feature relative abundance or presence-absence. The UniFrac metric was developed specifically for marker-gene survey data and incorporates feature phylogenetic relatedness by comparing the branch lengths for 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 relationship between features. Bray-Curtis and Jaccard dissimilarity index are example weighted and unweighted taxonomic metrics respectively (Bray and Curtis 1957; Jaccard 1912). These four groups of beta diversity metrics measure different community characteristics, and therefore they should not be used interchangeably but should be evaluated in a complementary manner to gain maximal insight into community differences (Anderson et al. 2011).

Previous studies have evaluated the impact of different bioinformatics pipelines (Sinha et al. 2017) and normalization methods (McMurdie and Holmes 2014; Weiss et al. 2017) on beta diversity metrics. Yet, the ability of these pipelines and normalization methods to account for sequence quality and coverage, and how this impacts beta diversity, remains unknown. Here we assess the effect of sequence characteristics on beta diversity calculations when data is processed using different bioinformatic pipelines and normalization methods. We employ a novel dataset consisting of mixtures of DNA extracted from stool samples with multiple technical PCR replicates, allowing us to evaluate (1) beta-diversity repeatability, (2) differences in beta diversity between individuals and treatments, and (3) ability to distinguish between groups of samples with varying levels of similarity. Furthermore, the data was produced from four replicate sequencing 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 (Olson et al. *in prep*). DNA extracts from stool collected from individuals (biological replicates) before and after exposure to pathogenic *Escherichia coli*. 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. In order to assess the measurement process technical variability technical replicates were generated at multiple levels, 16S rRNA PCR, sequence library generation, and sequencing run. Sequencing libraries were prepared at independent laboratories using the same protocol (ILLUMINA) with the sample 16S PCR as input, the resulting libraries were sequenced twice at each laboratory. Resulting

in four sequence datasets with varying sequence quality and library size variability. For the first laboratory (JHU) the base quality scores were lower than expected and the instrument was re-calibrated before the second run resulting in improved quality scores. For the second laboratory (NIST) the total run throughput was lower than expected, the pool library was re-optimized for resulting in increased throughput and lower sample to sample read count variability. Sequence data characterization was performed using the savR (Calder 2015) and ShortRead Bioconductor R packages (Morgan et al. 2009).

3.1. **Bioinformatic Pipelines.** Data from the four sequencing runs was processed using 6 bioinformatic pipelines including the QIIME open reference, closed reference, de novo, and deblur pipelines, as well as the Mothur de novo and DADA2 sequence inference pipelines. Code used to run the bioinformatic pipelines is available at https://github.com/nate-d-olson/mgtst_pipelines/, on the multirun branch. The Mothur pipeline uses the OptiClust algorithm for de novo clustering (Westcott and Schloss 2017). Preprocessing 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 Mothur implementation of the RDP bayesian classifier (Wang et al. 2007). 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. The forward and reverse reads were trimmed to 260 and 200 bp respectively. Using the values from the online protocol resulted in total abundance values around 5000. Forward and reverse primers were trimmed using cutadapt version 1.14 (https://cutadapt.readthedocs.io/en/stable/) (Martin 2011). DADA2 version 1.6.0 (B. J. Callahan et al. 2016) and reference database info. 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 (B. Callahan et al. 2016) using the DECIPHER R package version for multiple sequence alignment (Wright 2016) and the phangorn R package for tree construction (Schliep et al. 2017). For the QIIME pipelines all used the same input merged paired-end, quality filtered set of sequences (Caporaso et al. 2010). Both open and closed reference pipelines used the Greengenes 97% similiarity database for reference clustering. 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). The phylogenetic tree was constructed using FastTree and a multiple sequence alignment generated using pyNAST and the Greengenes reference alignment (Greengenes info) (Caporaso et al. 2010; Price, Dehal, and Arkin 2010). Additionally, sequence variants were inferred from the QIIME merged and quality filtered sequences using the Deblur sequence inference clustering method (version 1.0.3) (Amir et al. 2017). The same taxonomic classification and phylogenetic tree construction methods used for the other QIIME pipelines were also used for the Deblur clustered sequence data.

3.2. **Normalization Methods and Beta-Diversity Metrics.** Normalization methods are used to account for differences in sampling depth, number of sequences generated per sample, across samples. Rarefaction, subsampling counts without replacement to an even abundance is a commonly used method in macro-ecology and 16S rRNA marker-gene surveys (Gotelli and Colwell 2001; Hughes and Hellmann 2005). Samples were rarified to four level; 2000, 5000, and 10000 total abundance per sample, and to the total abundance of the 15th percentile. Rarefaction levels were selected based on values commonly used in published studies (Thompson et al. 2017), other comparison studies (Weiss et al. 2017; McMurdie and Holmes 2014). Rarified count data was analyzed using both weighted and unweighted Beta-diversity metrics. Other normalization methods were only analyzed for weighted metrics as these methods would not impact unweighted metric results. Other normalization methods include those previously developed for normalizing microarray and RNAseq data that are commonly used to normalize 16S rRNA marker-gene survey including upper quartile (UQ), trimmed mean of M values (TMM), and relative log expression (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Cumulative sum scaling (CSS) (Paulson et al. 2013) a normalization method developed specifically for 16S rRNA marker-gene survey data and total sum scaling (proportions, TSS) were also included in our weighted Beta-diversity metric assessment.

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 our feature-level Beta-diversity assessment the Bray-Curtis weighted and Jaccard unweighted metrics were used (Bray and Curtis 1957; Jaccard 1912).

3.3. **Beta-Diversity Assessment.** Our assessment consisted of three components, (1) beta-diversity repeatability, (2) differences in beta diversity between individuals and treatments, and (3) ability to distinguish between groups of samples with varying levels of similarity. Standard linear models were used to test for significance using the R `lm` function. Mixed effects models were used to take into account repeated measures were fit using the R `lmer` 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.

3.3.1. *Beta-Diversity Repeatability.*

- To assess beta-diversity repeatability we compared beta-diversity values between PCR replicates within the four sequencing runs.
- Sequencing runs different characteristics, variance in sample total abundance and sequence quality
- Compared mean beta diversity between PCR replicates for across pipelines for the four beta diversity metrics.
- Linear model used to quantify differences between pipelines and across sequencing runs for each of the diversity metrics calculated for the raw count data.
  - R model equation `mean_dist~pipe*seq_run_id`, `mean_dist` is the mean beta-diversity between the four PCR replicates, `pipe` is bioinformatic pipeline, and `seq_run_id` is sequencing run number (note does not account for lab effect).
- Evaluated the impact of normalization methods for NIST run 1.

- Linear model used to quantify normalization method impact on PCR repeatability.
  - Independent models fit for each diversity metric-pipeline combination.
  - R model equation `mean_dist~normalization`.

3.3.2. *Beta Diversity Signal to Noise Ratio.*

- Relationship between beta diversity between titrations (signal) and PCR replicates.

- This assessment evaluated the ability to differentiate titrations and post-exposure samples from pre-exposure samples.

- Signal was measured as the median beta diversity between pre-exposure PCR replicates and PCR replicates for each titration and post-exposure samples.
- Noise was measured as the mean of the median beta diversity between pre-exposure PCR replicates and median beta diversity between PCR replicates for the samples being compared to the pre-exposure samples.
- A weighted average of the signal to noise was calculated as the area under the curve (using the `trapz` function) of the signal/noise ratio and the proportion of pre-exposure sample in the in titration being compared (Borchers 2018).

- A linear model was used to quantify the differences in the signal to noise ratio between sequencing runs for each bioinformatic pipeline and diversity metric `log10(auc) ~ pipe + seq_run_num + biosample_id`.
- A mixed effects model was used to quantify the impact of different normalization methods on the signal to noise ratio for NIST run 1, `log10(auc) ~ method + (1 | biosample_id)`.
- Independent linear models were fit for each bioinformatic pipeline and diversity metric.

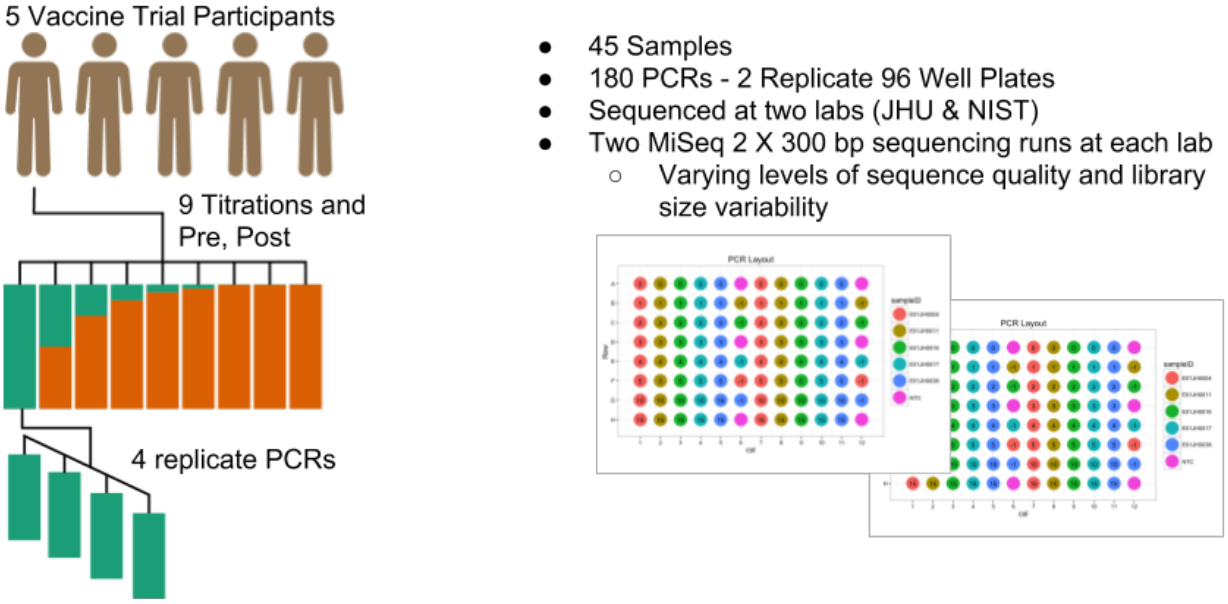

FIGURE 1. Stand-in figure.

3.3.3. *Beta Diversity Between Individuals and Treatments.*

- To quantify the contribution of biological and technical variability to total variability the distribution of beta diversity dissimilarity metrics were compared between individuals, within individual between conditions (pre- and post-exposure), and different types of technical replicates.

- Linear model used to quantify differences in beta diversity between biological and technical sources of variability.
  - R model equation `log(value) ~ variation + variation_label`, `value` is the beta diversity between replicates, `variation_label` is the variation source, and `variation` is the type of variation.
  - Variation labels
    - btw_subject_w/in_time: between subjects within treatment (pre- and post-exposure), within sequencing run
    - w/in_subj_btw_time: within subject between treatments, within sequencing run
    - btw_time: across subjects between treatments, within sequencing run
    - w/in_lab_pcr: with subject, treatment and lab, between sequencing runs?? (check)
    - w/in_lab_runs: within subject, treatment, lab, and sequencing runs (4 pcr replicates)
    - btw_labs: within subject, treatment, across sequencing labs
- Used variation partitioning (Borcard, Legendre, and Drapeau 1992) to quantify how technical and biological factors contribute to the total observed variation.
- Variation partition was calculated using the Vegan R package (Oksanen et al. 2018).
- Tested whether sources of variation significantly contributed to the overall variation using `drda` and `anova` (Oksanen et al. 2018).

4. RESULTS

The beta diversity assessment framework includes three components; (1) evaluating beta diversity between PCR replicates for sequencing runs with different error rates and library sizes, (2) difference in beta diversity between biological and technical replicates, and (3) ability to differentiate between PCR replicates from unmixed pre-exposure samples and post-exposure samples as well as the unmixed post-exposure samples.

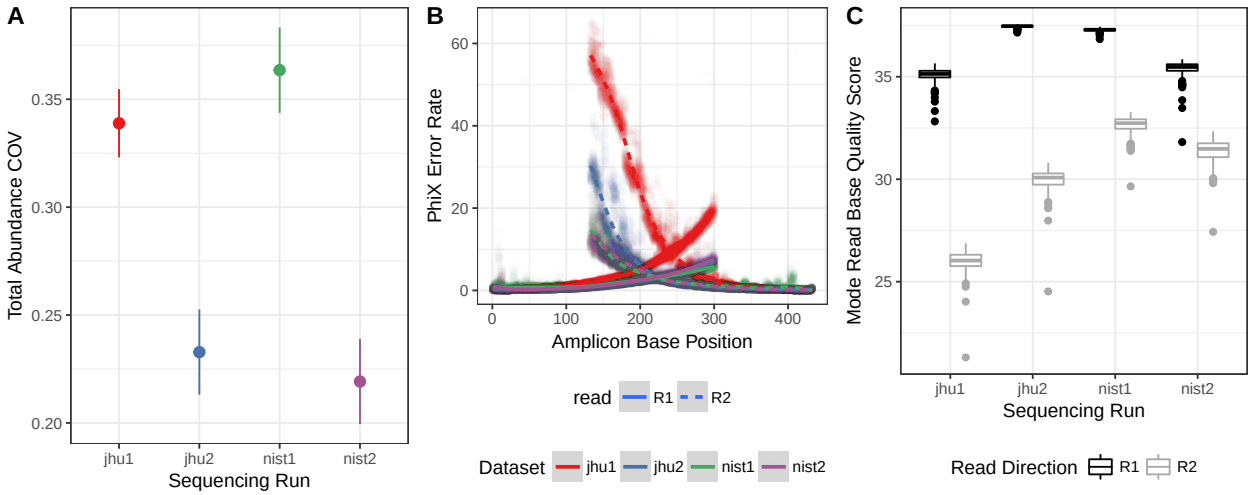

FIGURE 2. Sequencing quality and sample total abundance variation for the four sequencing runs used in this study. The same set of 192 PCRs were sequenced in all four runs. Independent sequencing libraries were generated at the two sequencing laboratories (JHU and NIST). (A) Sequencing run total abundance coeffient of variation estimate and 95% confidence interval calculated using a mixed effects linear model. (B) PhiX error rate relative to 16S rRNA amplicon base position for the four sequencing runs. (C) Distribution of mode read quality score by sequencing run.

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 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)

For our assessment we used a dataset consisting of two-sample titrations of DNA extracts from five vaccine trial participant stool samples collected before and after exposure to pathogenic *E. coli* (Fig. 1).

4.1. **Dataset Characteristics.** Sequencing data for the two-sample titration dataset was sequenced twice at two different laboraties, for technical replicate sequencing libraries and within laboratory sequencing runs. The four replicate sequencing runs varied sequence quality and feature total abundance variability (Fig. 2, *Supplemental DADA2 qual plot*). For both sequencing laboratories (JHU and NIST) the first runs had higher variability in feature total abundance between samples than the second runs (Fig. 2A). Sequencing error rate and base quality scores also varied by sequencing run. The first JHU run had higher PhiX error rates compared to the other sequencing runs especially for the reverse reads (Fig. 2B). Read base quality was lower for the reverse read than the forward reads for all four sequencing runs (Fig. 2C). Sequence data from the two NIST runs had higher quality scores compared to the JHU runs, excluding JHU2 forward reads (Fig.

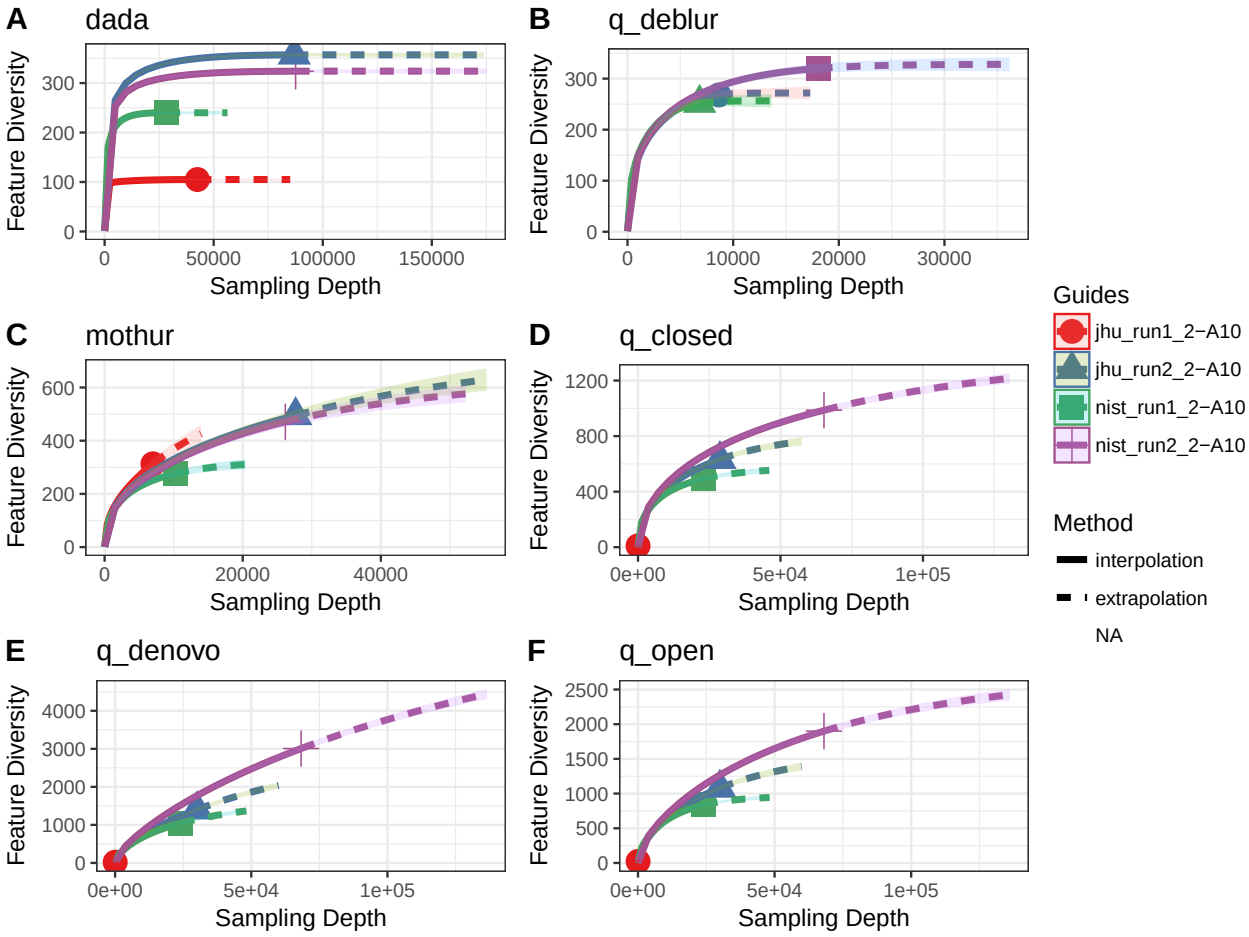

FIGURE 3. Rarefaction curves for an example sample across pipelines (A-F) and sequencing runs (line color).

2C). Overall JHU run 1 had low read quality and high total abundance COV, NIST run 1 had higher quality and total abundance COV, whereas the NIST and JHU second runs had lower COV and higher quality. By comparing the first JHU run results to the other runs we can evaluate how well the bioinformatic pipelines handle low quality reads. Similarly we can use data from the NIST run 1 to evaluate how well normalization methods are able to account for differences in total abundance between samples.

The six bioinformatic pipelines evaluated in the study employed different pre-processing, clustering, and quality filtering 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, proportion of sequences passing quality control (Table 1).

Sequence inference methods have lower species diversity estimates and reach asympotote, whereas *de novo*, open-reference, and closed-reference methods do not (Fig. 3). Based on the rarefaction curve slopes the QIIME *de novo* pipeline had the highest rate of artifacts, due to not filtering singletons (Fig. 3E). The sequence inference methods, DADA2 and Deblur plateau around the same level (Fig. 3A & B). However, DADA2 asymptotes were inconsistent across sequencing runs, indicating artificial plateaus for the lower throughput and lower quality runs (Fig. 3A). Mothur and Deblur rarefaction curves were consistent across sequencing runs. The QIIME open reference, closed reference, and *de novo* rarefaction curves were influenced by both sequence quality and library size (Fig. 3D-F).

4.2. **PCR Repeatability.** We evaluated differences in beta diversity between PCR replicates across sequncing runs to assess bioinformatic pipeline and normalization methods robustness to low quality sequence data

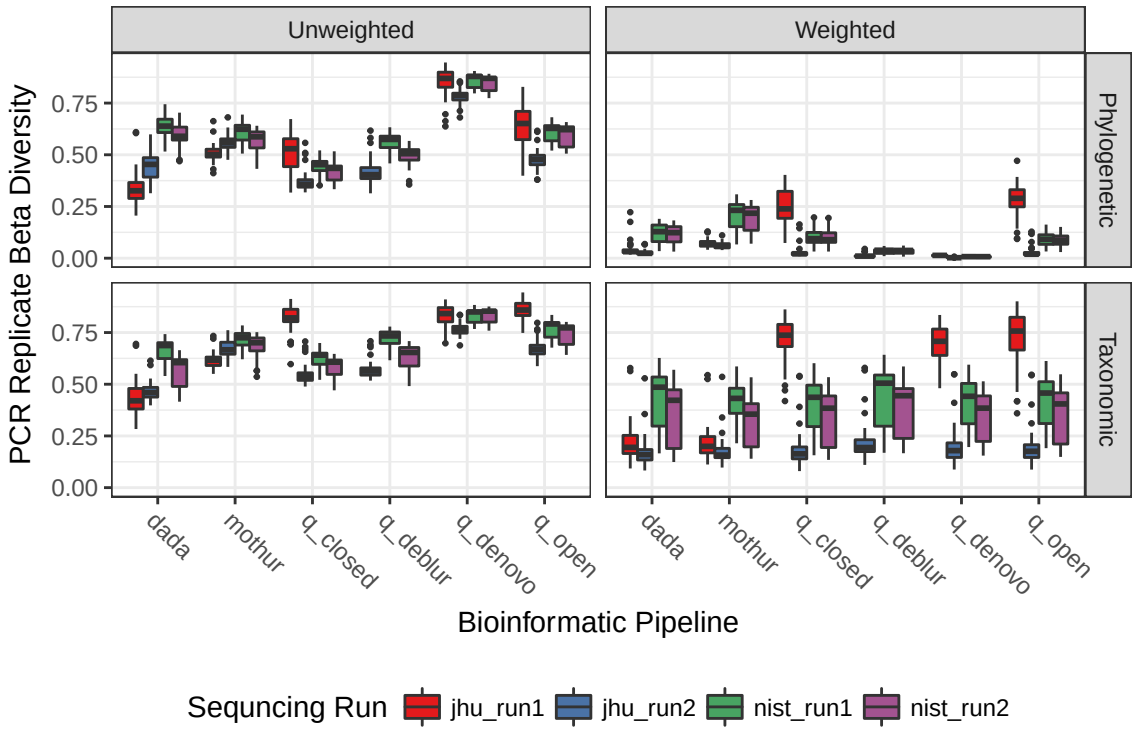

FIGURE 4. Distribution of mean pairwise PCR replicate beta diversity for by sequencing run and pipeline for un-normalized count data.

and variability in sample total feature abundance. Higher pairwise distances for NIST run 1 indicates diversity metrics were negatively impacted by larger variation in total abundance across PCR replicates (Fig. 2A). Whereas, higher pairwise distances for JHU run 1 indicates bioinformatic pipelines are negatively impacted by sequencing errors (Fig. 2B & C). The NIST and JHU second runs had low sequencing error rates sample total abundance variation and therefore we would expect the beta diversity between PCR replicates for these runs to be lower than the first runs.

We assessed the sequence quality impact on sequence quality on PCR replicate beta diversity using the un-normalized count data. Mothur and DADA2 mean PCR replicate beta diversity was consistent across the JHU runs, suggesting the pipelines are more robust to sequencing errors than the other pipelines evaluated in this study (Fig. 4). Conversely, the Deblur pipeline had the highest number of failed samples for the first JHU run, and therefore less robust to sequencing errors compared to the other pipelines (Table 1). Based on the sequencing run error rates and total abundance variation we expected the second NIST run to have comparable, if not lower, beta diversity between PCR replicates. However, the beta diversity was consistently higher for both the NIST runs than the second JHU run and the second NIST run values were more similar to the first NIST run than the second JHU run. PCR replicate beta diversity varied by diversity metric (Fig. 4). Beta diversity was consistently higher for unweighted compared to weighted metrics and phylogenetic diversity metrics were lower than taxonomic metrics.

Normalization methods varied in their ability to improve beta diversity repeatability. To evaluate the effect of normalizion methods on beta diversity repeatability we compared normalized to un-normalized PCR replicate beta diversity (Fig. 2). Most normalization methods had minimal impact on beta diversity repeatability. However, for a number of pipelines TMM and RLE normalization methods significantly lowered weighted PCR replicate beta diversity (Fig. 5A). In general the normalized count beta diversity was not significantly different from the unnormalized, aside from a few pipeline metric combinations. The lack of consistent effect on weighted beta diversity metrics for the other normalization methods indicates that normalization methods do not effect PCR replicate beta diversity or that normalization method effect is metric and pipeline specific.

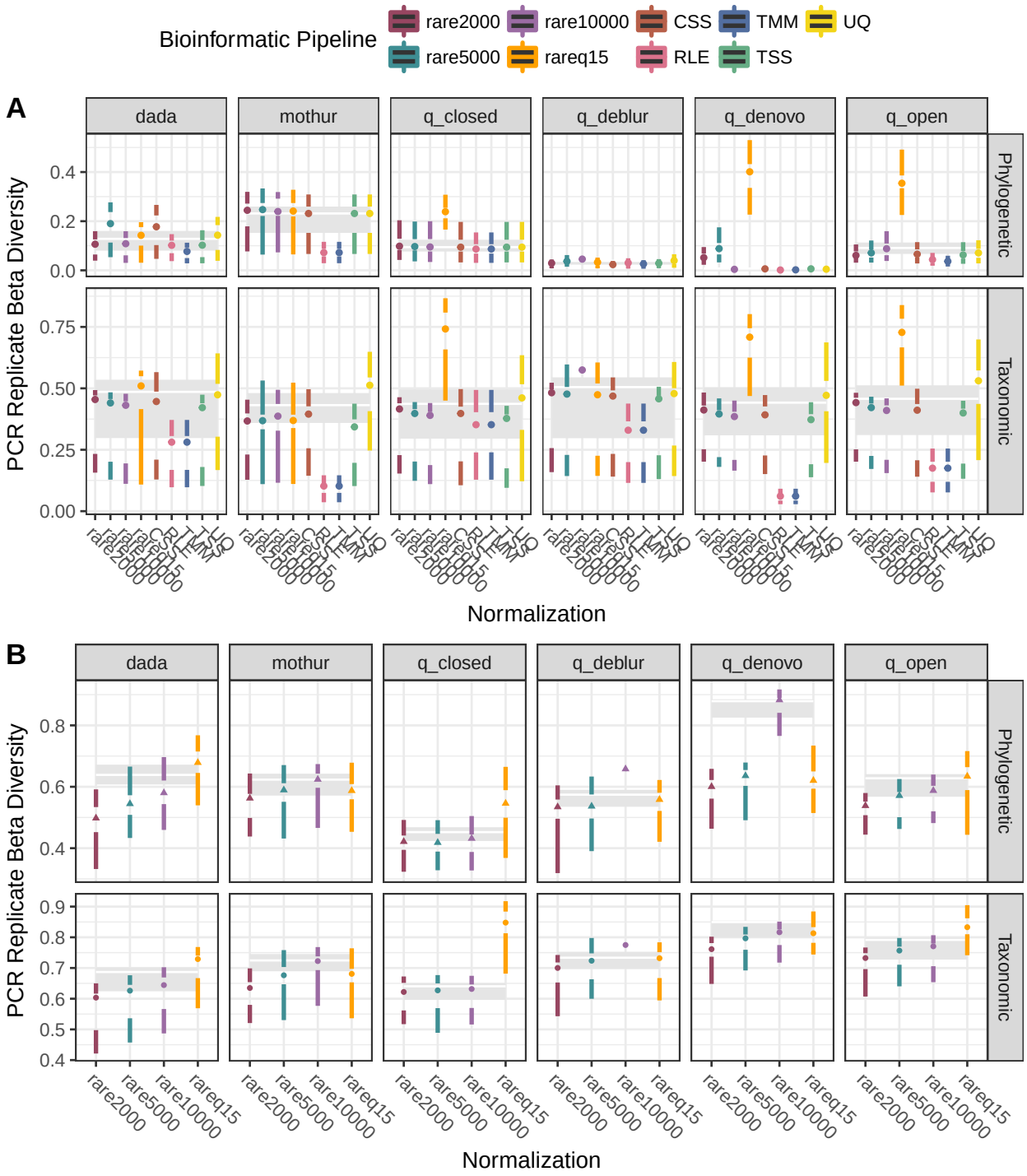

FIGURE 5. Impact of normalization method on mean weighted (A) and unweighted (B) pcr replicates beta diversity, for sequencing run with higher quality and total abundance variability, NIST run 1. Data are presented as minimal-ink boxplots, where points indicate median value, gap between point and lines 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

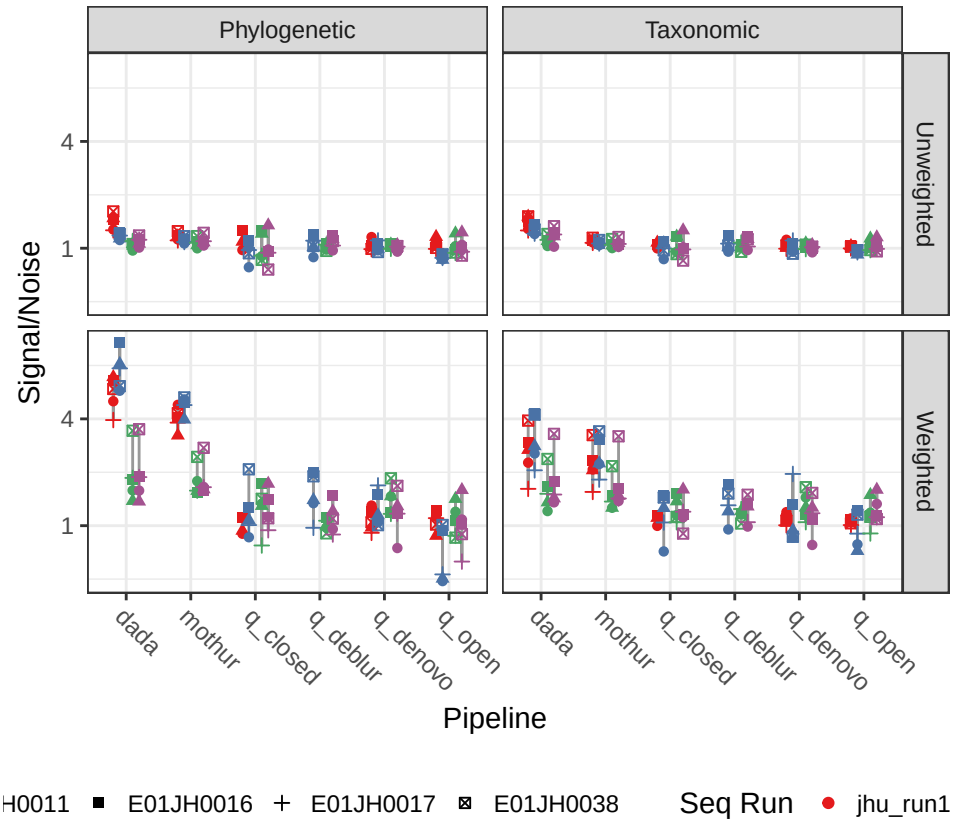

FIGURE 6. The weighted average signal to noise varied by pipeline, run, and diversity metric. Points indicate the signal to noise for each individual with grey lines representing the range of values for a pipeline and sequencing run.

PCR replicate beta diversity varied for count data normalized using rarefaction 2). Rarefying counts to the total abundance of the 85th most abundance sample (rareq15) resulted in no differences or lowered beta diversity repeatability for PCR replicates for all pipeline-diversity metric combinations except for mothur-jaccard and *de novo*-UniFrac . Rarefying counts to 2000, improved results for all but closed reference, though the effect was not significant for Deblur. Rarefying to 5000 total abundance produced similar results to 2000, though the effect was statistically significant for fewer pipeline-metric combinations. Overall, count data normalized using TMM and RLE had the lowest beta diversity between PCR replicates for weighted metrics, followed by TSS. For unweighted metrics, rarefying count data to 2000 total abundance resulted in the lowest beta diversity between PCR replicates.

4.3. **Signal to Noise.** Signal-to-noise ratio for unweighted metrics was around 1 for all pipelines and sequencing runs (Fig. 6). Signal-to-noise ratios around 1 indicate that the signal magnitude (biological differences) was comparable to the noise (differences between PCR replicates). We define the signal-to-noise ratio as the beta diversity between unmixed pre-exposure samples and the other samples in the titration series (signal) divided by PCR replicate beta diversity for the samples being compared. Only DADA2 and mothur signal to noise ratios were greater than the pipelines pre-processed using QIIME. DADA2 and Mothur higher signal to noise ratio differences for JHU runs compared to NIST runs, especially for weighted phylogenetic. The relationship between NIST and JHU runs for the signal to noise relationship is consistent with the PCR replicate beta diversity results.

Similar to PCR beta diversity repeatability we evaluated the how well normalization methods account for differences in the sample total abundance by comparing un-normalized to normalized NIST run 1 signal-to-noise ratio (Fig. 7). Normalizing count data should increase the signal-to-noise ratio. Most normalization methods did not have a significant effect on the signal-to-noise ratio for weighted metrics (Fig. 7A). TSS

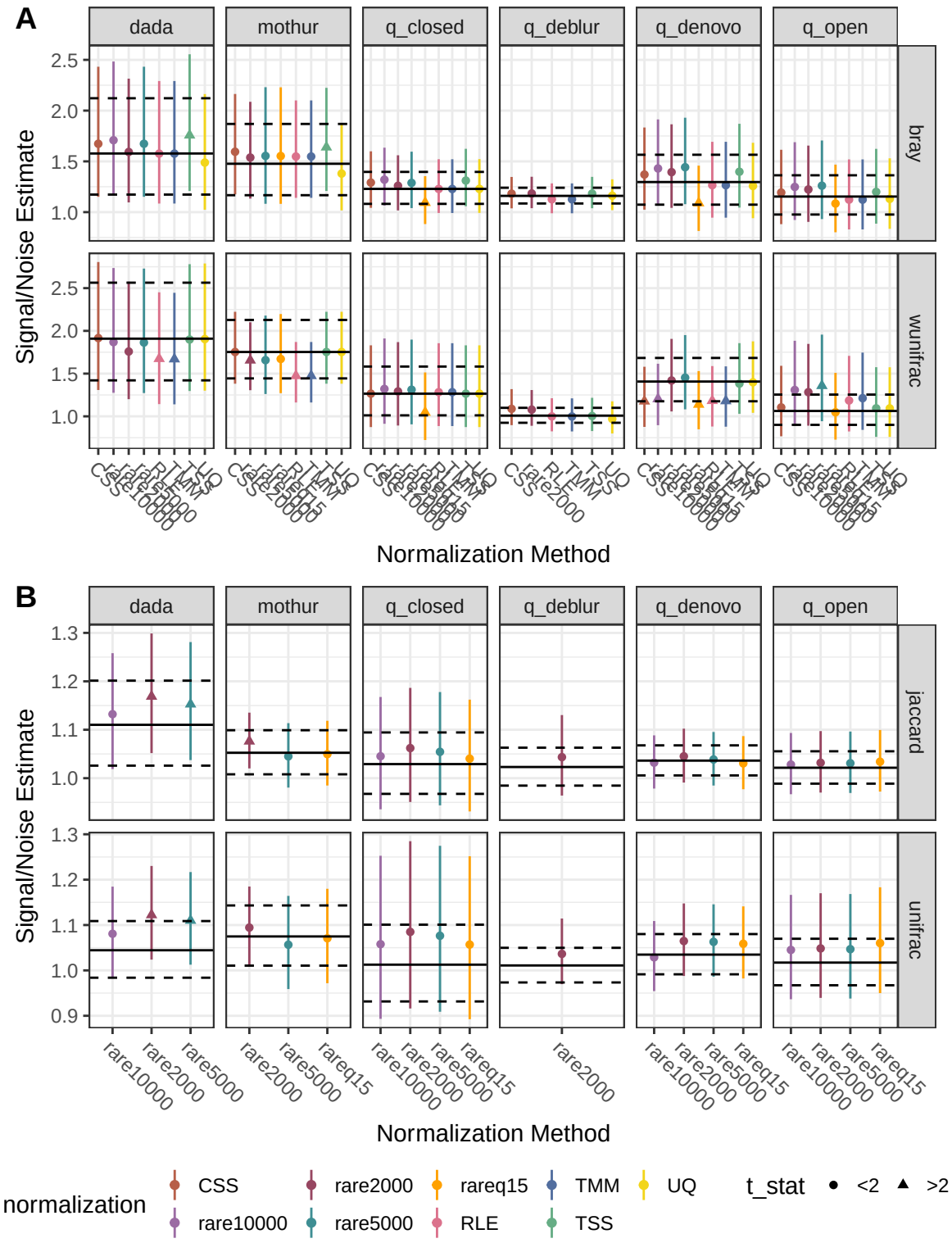

FIGURE 7. Weighted average signal to noise ratio estimate and 95 CI for raw and normalized count data for (A) weighted and (B) unweighted beta diversity metrics. Estimates calculated using a mixed effects linear model using subject as random effect. The horizontal solid line is the unnormalized count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

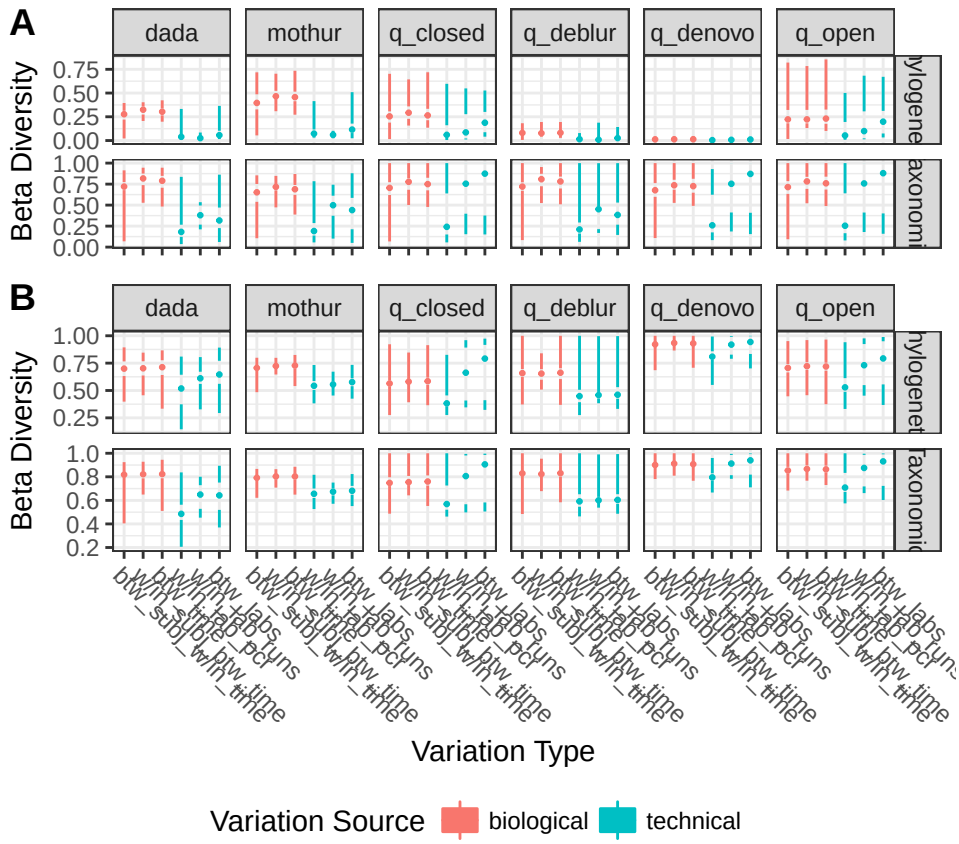

FIGURE 8. Biological vs. Technical Variation, distribution is beta diversity between technical replicates and biological treatments (subject and timepoint).

significantly increased the Bray-Curtis signal to noise ratio for Mothur and DADA2, but not for the other pipelines or weighted UniFrac. Some normalization methods significantly decreased the signal to noise ratio for weighted beta diversity metrics. Rarefying counts to 15th quantile resulted in significantly lower weighted UniFrac and Bray-Curtis beta-diversity for QIIME closed-reference and *de novo*. 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*. While, rareifying count data often increased unweighted metric signal-to-noise ratio though the increase was only significant for DADA2 and Mothur. Rarefying count data to 2000 and 5000 significantly improved the DADA2 signal-to-noise ratio for both unweighted metrics. Mothur signal-to-noise ratio was only significantly higher for Jaccard and when count data was rareifed to 2000.

4.4. **Biological v. Technical Variation.** We next looked at how different pipelines and normalization methods captured diversity differences between biological and technical replicates. Beta-diversity distances between biological and technical replicates varies by pipeline and beta-diversity metric for raw count data (Fig. 8). Overall, as expected, the mean diversity observed between biological replicates was greater than that between technical replicates. The magnitude of the difference varied by diversity metric and pipeline. The difference in beta diversity between biological replicates and technical replicates was greater for DADA2, mothur, and Deblur, compared to QIIME open-reference, close-reference, and *de novo* pipelines. The difference was greater for weighted than unweighted beta diversity metrics for all pipelines.

We also used variation partitioning to evalute the impact of different normalization methods on the amount of variation attributable to subject, titration factor (unmixed pre-exposure and unmixed post-exposure), and sequencing run. Accross all pipelines and diversity metrics, the greatest amount of variation is often explained by subject, followed by titration factor (Fig. 9). In our unnormalized pipelines, sequencing run accounts

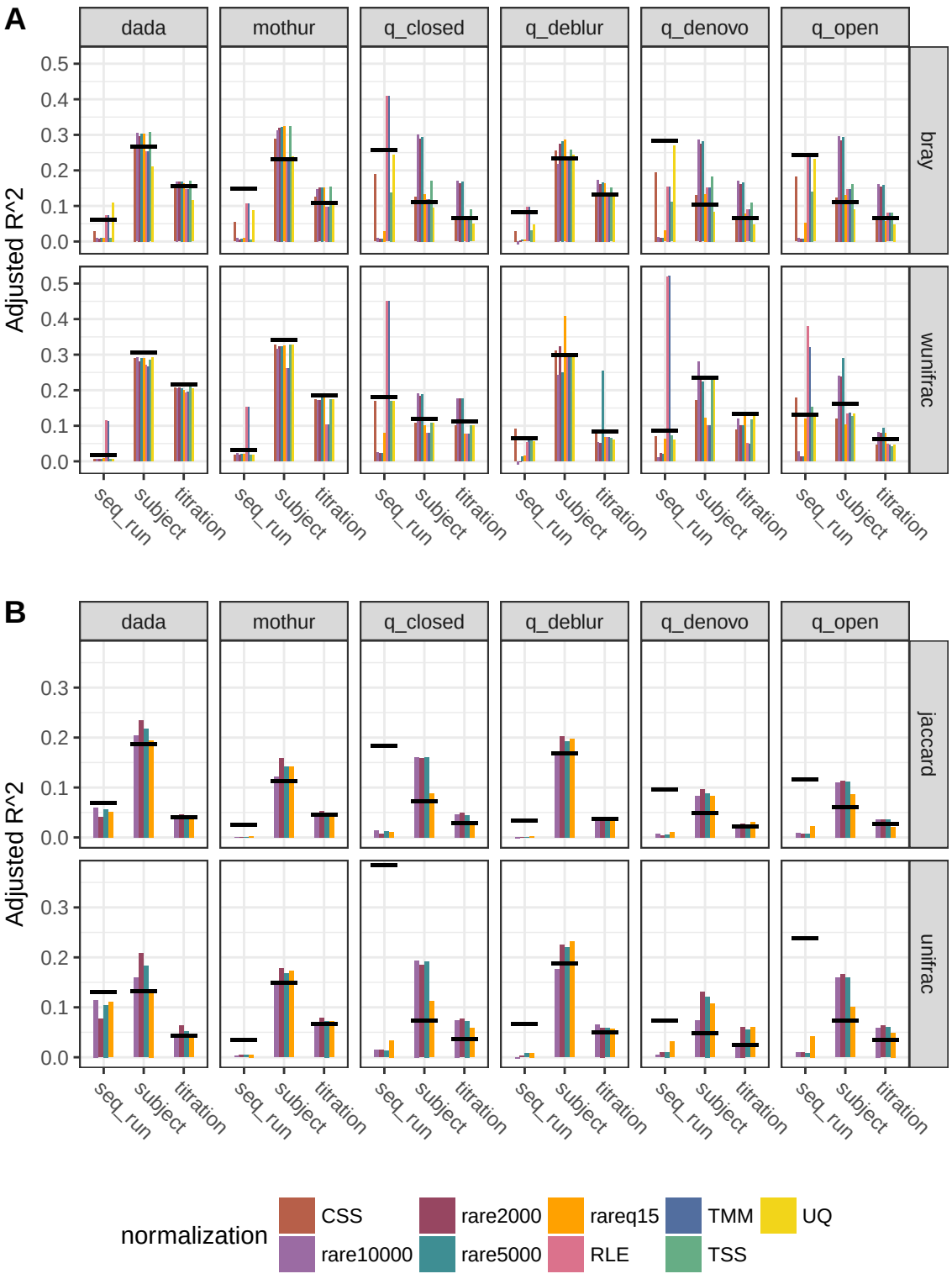

FIGURE 9. Biological vs. Technical Variation, y-axis is the adjusted R^2 value, indicating proportion of variance explained by each biological (subject and titration) and technical (seq run) variable. Black bars represent values for un-normalized data.

for a greater proprotion of the explained variance, highlighting the overall importance of normalizing our datasets. Effective normalization methods decrease the technical variability in the data without decreasing the biological variability. For both weighted and unweighted metrics rarefaction normalization methods generally show increased amounts of variation explained by biological factors rather than technical artifacts. The non-rarefaction normalization methods do not reduce the impacts of technical arifacts as effectively, especially for the QIIME pipelines. RLE and TMM consistently increase the technical variability and often decrease the amount of variability in the data attributed to biological factors (Fig. 8A). For the QIIME closed reference, open reference, and *de novo* pipelines and taxonomic beta diversity metrics subject variation greater than sequencing run after normalization (excluding TMM, RLE, UQ and CSS) but not without.

5. Discussion

Sequence data characteristics, specifically sequence error rate and variation in library size can negatively impact beta diversity analysis (**REF**). Bioinformatic pipelines are used to convert the raw sequence data to a count matrix, ideally differentiating true biological sequences from sequencing artifacts alleviating biases due to sequencing errors. Normalization methods, such as rarefying count data and cumulative sum scaling, are used to account for library size differences. For our assessment we compared the performance of six bioinformatic pipelines and nine normalization methods for four beta-diversity metrics. The results from our study employing a novel dataset and assessment framework indicate that bioinformatic pipelines vary in their ability to alleviate biases in beta diversity due to sequencing errors and some normalization methods accentuate the biases in beta diversity analysis due to library size differences.

Our assessment framework consisted of three components, (1) PCR beta-diversity metric repeatability, (2) signal-to-noise analysis, (3) difference in variation between biological factors and technical replicates. Our assessment framework utilized a novel two-sample titration dataset with multiple levels of technical replication including 16S rRNA PCR, sequencing library, and sequencing run. Multiple PCR replicates were used to assess beta-diversity repeatability. The titrations were used to assess signal-to-noise ratios. Finally, the different sample types (trial participants and exposure status) in conjunction with multiple sequencing runs were used to compare sources of variability.

Using PCR replicate beta diversity for the four sequencing runs we showed how sequence quality and total abundance variation impacts PCR replicate beta diversity repeatability was pipeline and diversity metric dependent. For the QIIME *de novo* pipeline the PCR replicate unweighted UniFrac was high for all runs but low for weighted UniFrac. We attributed the difference between QIIME *de novo* the weighted and unweighted UniFrac metric to the high number of singletons in the dataset. These sigletons indicate the in ability of the pipeline to group sequencing artifacts with true biological sequences. Singleton removal, a step included in the QIIME open-reference pipeline addresses this bias. Sequence data from JHU run 1, which had lower error rate and total abundance relative to the other sequencing runs had consistently better repeatability across pipelines and diversity metrics. Normalization methods improved repeatability, excluding rarefying data to 15th quantile, which decreased repeatability especially for QIIME pipelines. TMM improved weighted beta diversity repeatability for NIST datasets, greater variability in library size. While it is important to reduce the beta-diversity between technical replicates, it is more important to be able to detect true differences between biological samples. To detect differences between biological samples, sample dissimilarity due to biological factors must be greater than sample dissimilarity due to technical variability or noise.

For the second component of our assessment we compare the signal-to-noise ratio between sequence data characteristics, bioinformatic pipeline, and normalization method. The assessment results varied by pipeline and diversity metric, with DADA2 and mothur consistently out performing the other bioinformatic pipelines for weighted metrics. **MORE**

To evaluate the impact of difference sources of variability on bioinformatic pipeline and normalization methods we compared the beta diversity between different biological samples and technical replicates, including PCR replicates and sequencing runs. Differences in beta diversity between biological samples and technical replicates varied by diversity metric, pipeline, and normalization method. Overall differences in beta diversity metrics are due to differences in how the four metrics measure community similarity. For phylogenetic metric, the beta diversity tended to be lower compared to the taxonomic metrics. This was due to low

overall phylogenetic diversity, or the majority of the features being phylogenetically closely related. For most pipelines and beta diversity metrics, normalizing the count data increased the difference in beta diversity between biological and technical replicates, resulting in a greater ability to detect beta diversity differences between communities from different treatments. However, some metrics, namely rarefying to the 15th quantile, RLE, and TMM, frequently reduced the difference between the biological signal and noise due to technical variability. Variation partitioning results were consistent with this conclusion. RLE and TMM were developed for normalizing microarray and and RNAseq data and not marker-gene sequence data. While these normalization methods have been show to be useful for differential abundance analysis, they are not appropriate for beta-diversity analysis.

5.1. **Conclusions.** The results presented in this study can be used to help determine the 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. These differences impacted the beta diversity repeatability. Normalization can help improve repeatability, but sometimes at the cost of decreasing the difference between biological signal and technical variability. Mothur and DADA2 are more robust to lower quality datasets. Normalization methods can improve ability to detect true biological signal though normalization methods developed for gene expression methods may not be appropriate.

Bioinformatic pipelines combine multiple algorithms to convert the raw sequence data into a count table for use in statistical analysis. The 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 optimizeing 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 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 EP. 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):2460–1.

Edgar, Robert C, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. 2011. "UCHIME Improves Sensitivity and Speed of Chimera Detection." *Bioinformatics* 27 (16):2194–2200.

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):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." *Nat. Methods* 10 (12):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." *Appl. Environ. Microbiol.* 75 (23):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." *Appl. Environ. Microbiol.* 73 (16):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.