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

1. 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.

2. Methods

2.1. **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 [**REF**].

2.2. **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.

2.3. **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).

2.4. **Beta-Diversity Assessment.**

- Three assessment 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.

2.4.1. *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.
- Used **XYZ model** to test for differences.
- Evaluated the impact of normalization methods for NIST run 1.
- Used **XYZ model** to test for significance.

2.4.2. *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

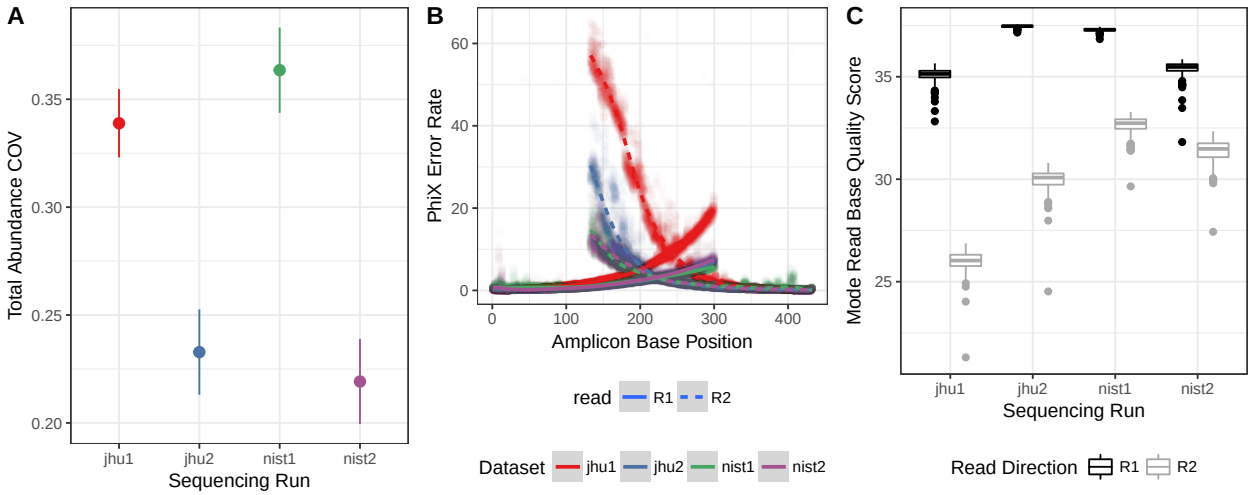

FIGURE 1. A. Coeffient of variation in total abundance by sequencing run estimate and 95% confidence interval obtined using a mixed effects linear model. B. PhiX error rate relative to 16S rRNA amplicon base position for initial and the four sequencing runs. C. Distribution of mode read quality score by sequencing run.

conditions (pre- and post-exposure), and different types of technical replicates.

- Used **XYZ model** to test for significance.
- Used variation partitioning (**REF**) to quantify how technical and biological factors contribute to the total observed variation.

2.4.3. *Differentiation Power.*

- We assessed the impact of bioinformatic pipeline and normalization methods on beta-diversity metrics using a similar approach used in (Reynolds et al. 2006; McMurdie and Holmes 2014).
- This assessment evaluated the ability to differentiate titrations and post-exposure samples from pre-exposure samples.
- To summarize performance across titration comparisons, area under the curve was calculated using the trapazoid method (**R package ref**).
- Use **XYZ model** to test for significance.

3. 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. 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* (**Experimental Design Figure**).

3.1. **Dataset Characteristics.** Bioinformatic pipelines and normalization methods are used reduce the impact of noise in marker gene sequencing data due to sequencing errors and differences in the library size between samples. Sequencing data for the two-sample titration dataset was obtained from four replicate

TABLE 1. Summary statistics for the different bioinformatic pipelines. Four pipelines, de novo, open reference, closed reference, and deblur (sequence inference), used the sample sequence pre-processing methods. DADA2 is a denoising sequence inference pipeline and mothur is a de-novo clustering pipeline. No template controls were excluded from summary statistics. Sparsity is the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2) in the count. Singletons is the total number of OTUs comprised of a single read in a single sample. Sample coverage is the median and range (minimum - maximum) per sample total feature abundance. Filter rate is the proportion of reads that were removed while processing the sequencing data for each 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)
q_deblur	3711	0	576	0.940	9135 (30423-4)	0.14 (0.24-0)
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_denovo	180834	120599	766	0.994	26250 (118767-4)	0.37 (0.75-0)
q_open	45663	39	766	0.981	26373 (118421-3)	0.37 (0.75-0)

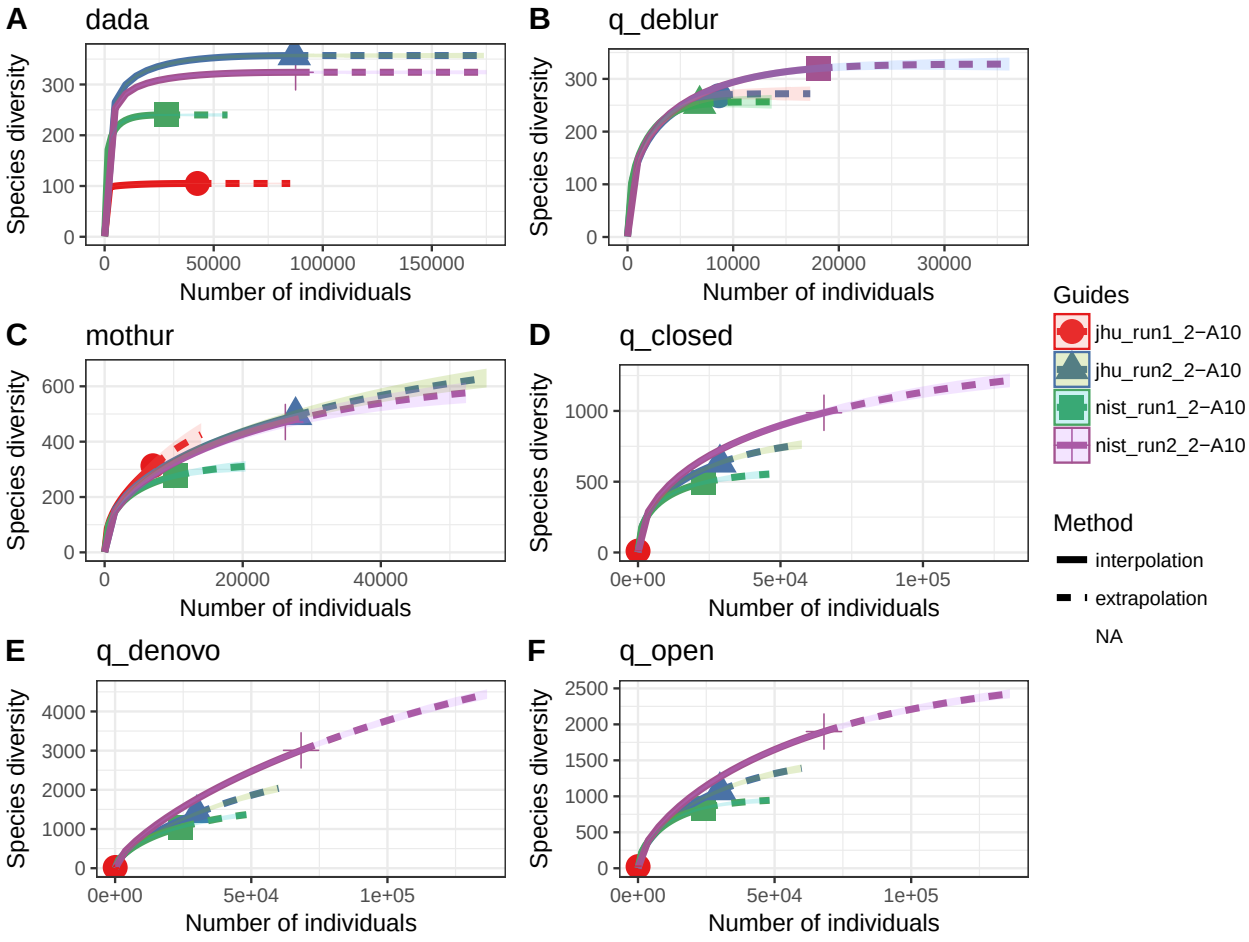

FIGURE 2. Rarefaction curves for example sample across pipelines and sequencing runs.

sequencing runs with different sequence quality and library size variability (Fig. 1, *Supplemental DADA2 qual plot*). The sequence data was processed using six different bioinformatic pipelines, DADA2, mothur, Deblur, and QIIME - *de novo*, open-reference, closed-reference. The NIST runs had but greater variabiltiy in library size (Fig. 1A). Good separation between sample and no template control library size for JHU but not NIST samples.
However, total abundance is lower for samples compared to no template controls for most sequencing runs and samples. Though a few no template controls have values within the sample range. (see pipe characterization total abundance and pass rate plots). After processing the sequence data with the six bioinformatic pipelines the coefficint of variation total abundance between PCR replicates was lower for JHU and NIST run 2 compared to the first runs (Fig. 1B).
The first JHU run had higher PhiX error rates compared to the other sequencing runs especially for the reverse reads (Fig. 1C). The read base quality was lower for the reverse read than the forward reads, the reverse read quality score was higher for the two nist runs compared to the JHU runs (Fig. 1C). The forward reads from NIST run 1 had the best read quality score followed by JHU run 2.
JHU run 2 had low read quality and high total abundance COV, NIST run 1 had higher quality and total abundance COV, NIST and JHU second runs had lower COV, but JHU had lower read 2 quality and NIST run had lower forward read quality. The differences in sequence quality and total abundance variabiltiy between sequencing runs allows us to evaluate how well bioinformatic pipelines and normalization methods handle low quality reads and variability between samples.

The six bioinformatic pipelines evaluated employ 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, and proportion of sequences passing quality control (Table 1). Rarefaction curves are in ecology to determine how well a community has been sampled (Gotelli and Colwell 2001,Chao et al. (2014)). Measurement methods prone to errors, such as marker-gene sequencing, will never reach the asympotote if errors are not appropriately accounted for in sample processing (Chiu and Chao 2016). Sequence inference methods have lower species diversity estimates and reach asympotote, whereas *de novo*, open-reference, and closed-reference methods do not (Fig. 2). Based on the rarefaction curve slopes the QIIME *de novo* pipeline had the highest rate of artifacts, due to not filtering singletons. The sequence inference methods, DADA2 and Deblur plateau around the same level. However, DADA2 asymptotes were inconsistent across sequencing runs, indicating artificial plateaus for the lower throughput and lower quality runs. 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.

3.2. **Technical Artifacts.** We evaluated differences in beta diversity between PCR replicates between the four sequncing runs to assess how robust the different bioinformatic pipelines and normalization methods are to low quality sequence data and variability in per sample total abundance. Higher pairwise distances for NIST runs indicate diversity metrics negatively impacted by larger variation in total abundance across PCR replicates. Whereas higher pairwise distances for JHU run 1 indicates bioinformatic pipelines are less robust to sequencing errors. Mean pairwise distance was greater for qiime De-novo, open and closed reference pipelines for JHU1 (lower quality and greater variability) relative to the other sequencing runs for all metrics excluding weighted UniFrac (Fig. 3). QIIME *de novo* had high mean pairwise distance across sequencing runs for unweighted UniFrac and low for weighted Unifrac. The low weighted UniFrac for QIIME *de novo* is potentially due to large number of singletons in the dataset, ~120K out of ~180K total features. These singletons are likely sequencing errors and therefore closely related to other taxa therefore minimally impact the weighted unifrac results.
DADA2 pairwise distances were greater for NIST1 and NIST2 (which had greater variability in library size) compared to the JHU2 run which had the lowest pairwise distance.
Mean pairwise distances were consistent across the JHU runs for Mothur and DADA2, suggesting they are better able to account for sequencing errors than other pipelines evaluated in this study. Conversely, the Deblur pipeline had the highest number of failed samples for JHU1, suggesting it is less robust to sequencing errors compared to the other pipelines (Table 1).

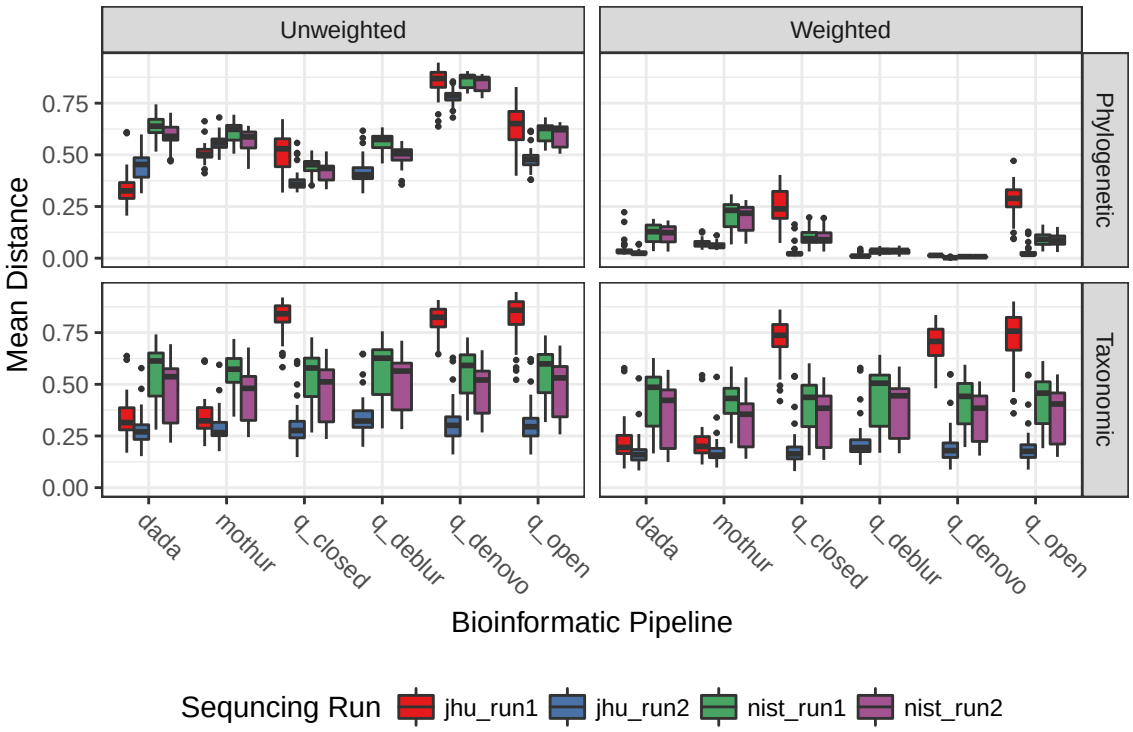

FIGURE 3. Distribution of mean pairwise beta diversity for PCR replicates by sequencing run and pipeline for raw count data.

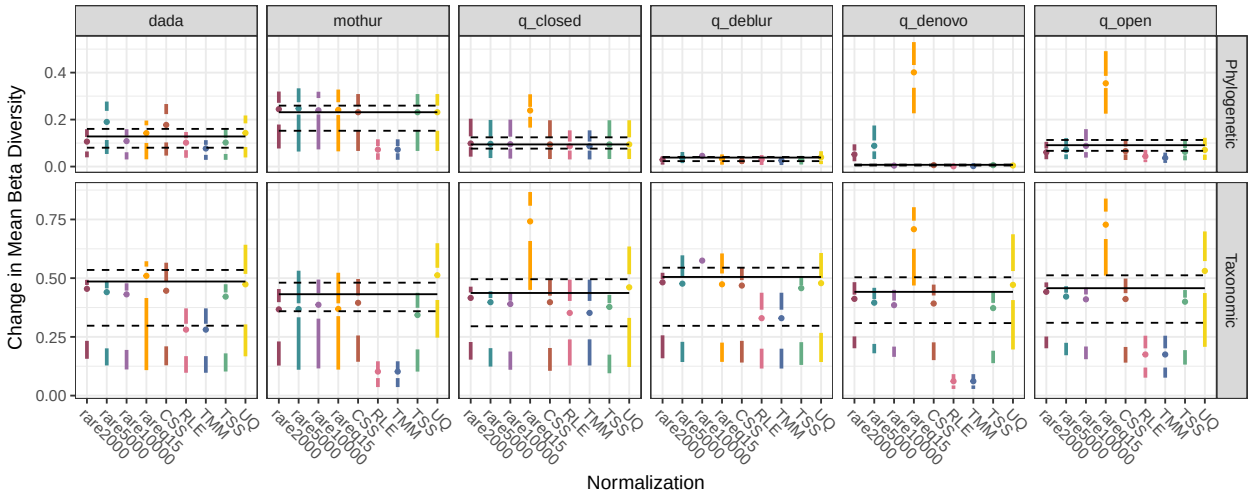

FIGURE 4. Impact of normalization method on mean weighted beta diversity between pcr replicates, for sequencing run with higher quality and total abundance variability, NIST run 1. Data are presented as minimal-ink boxplots, where points indicate median value, gap between point and lines the interquartile 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

3.3. **Biological v. Technical Variation.** We next looked at how different pipelines and normalization methods captured diversity differences between our biological and technical replicates. Beta-diversity distances between biological and technical replicates varies by pipeline and beta-diversity metric (Fig. 6). Overall, as expected, the mean diversity observed between biological replicates was greater than that between technical

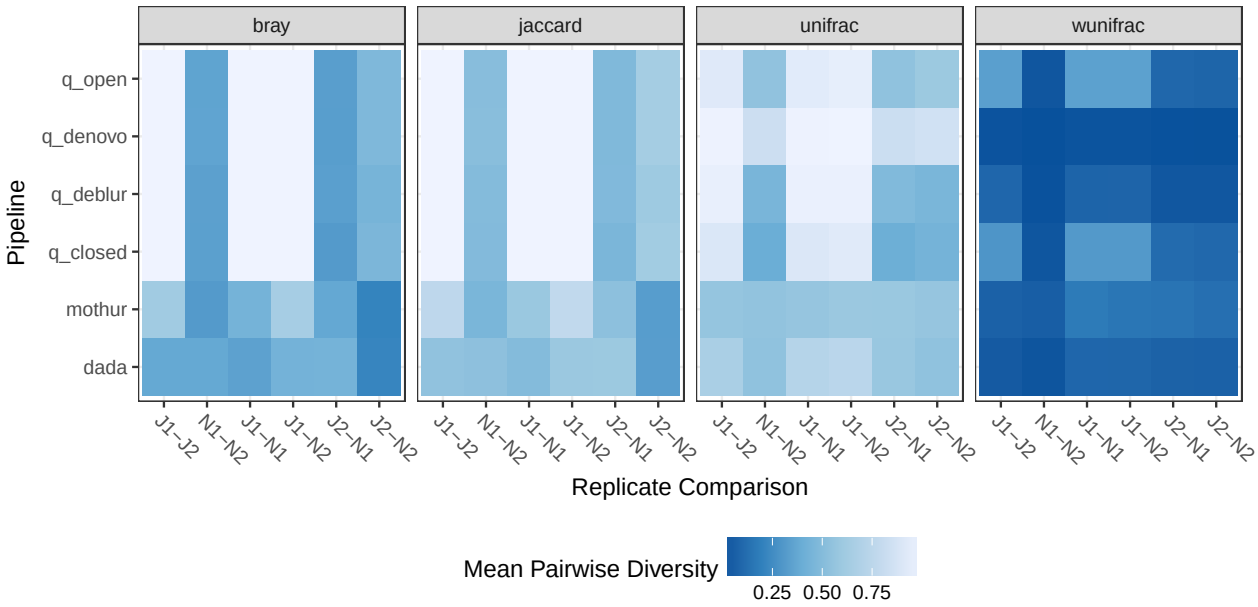

FIGURE 5. Heatmaps indicating mean beta diversity of un-normalized PCR replicates sequenced at the same lab (J1-J2 or N1-N2) or at different labs (J[1-2]-N[1-2]) for each pipline and metric.

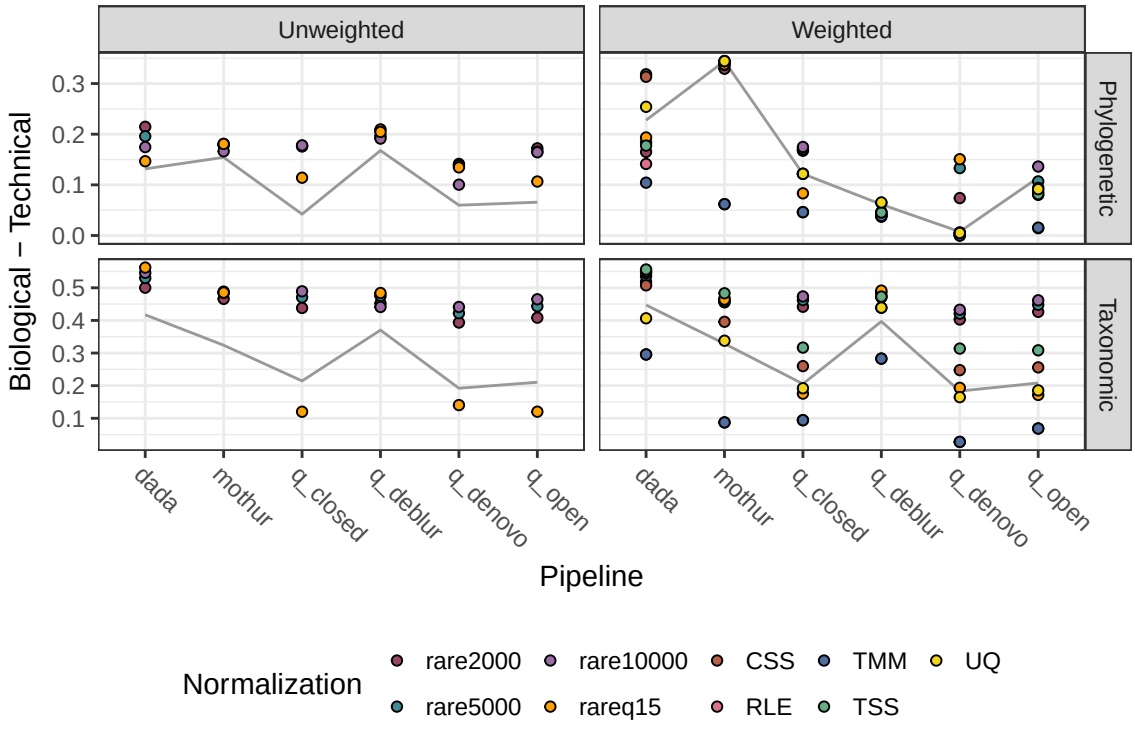

FIGURE 6. Biological vs. Technical Variation, y-axis is the differences between the mean biological (subject and titration level) and technical variation (sequencing lab and run) (pairwise distance between replicates.) Grey line indicates the mean differences for diversity metrics calculated using raw counts. Higher values indicate better differentiation between technical variability and true biological differences. Points indicate mean differences for diversity metrics calculated using normalized counts with color indicating normalization method.

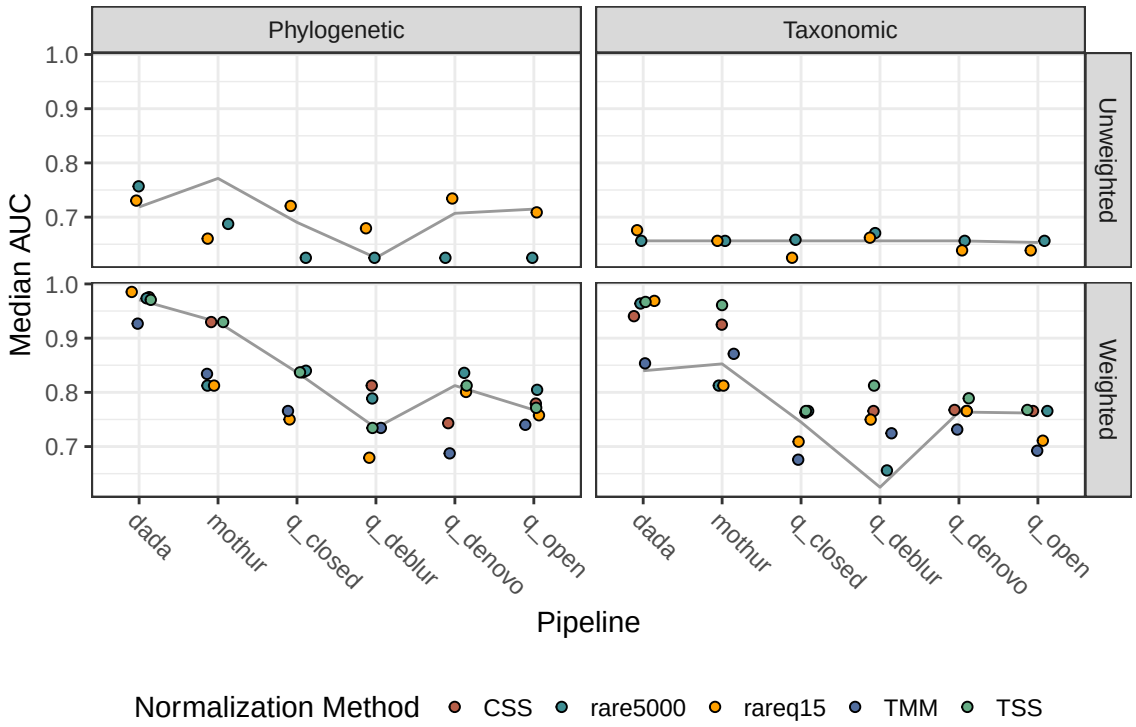

FIGURE 7. Comparison of median AUC for clustering results across pipelines and normalization methods for four beta diversity metrics. Grey line indicates, the median AUC for unnormalized, raw, count table values. Points above the grey line are normalization methods that improve performance and below are methods that decrease performance.

replicates. Generally, greater differences between biological and technical replicates were observed using taxonomic metrics, rather than phylogenetic metrics (with the exception of mothur for weighted phylogenetic metrics). For weighted metrics, TMM decreased the difference relative to raw counts indicating that for beta diversity analysis these normalization methods reduce the power to distingush true biological differences from technical variability or noise. For unweighted metrics, normalization by rarefaction produced consistent differences in variation accross most subsampling depths. The three QIIME pipelines, however, identified smaller differences between biological and technical replicates when subsampled to the 15th quantile. This inconsistency is most likely due to greater sample loss in these pipelines at this subsampling level [?? is this true].

We also used variation partitioning to determine 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. **??**). In our unnormalized pipelines, sequencing run accounts 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. 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. **??**).

3.4. **Comparison to Expectation.** Performance varied by pipeline with DADA2 having consistently higher performance compared to the other pipelines (Fig. 7).
Rarefaction level had inconsistent performance relative to unnormalized data. Rareified to the 15th quantile library size improved performance relative to unnormalized data with qiime pipelines when using UniFrac but lower performance for Jaccard.

For weighted metrics, normalization method performance relative to unnormalized counts varied by pipeline, though TMM and rarefaction to 15th quantile had consistently lower performance compared to unnormalized data.

4. Discussion

Sequence data characteristics, specifically sequence error rate and variation in library size can negatively impact beta diversity analysis. The bioinformatic pipeline used to convert the raw sequence data to a count matrix ideally differentiate true biological sequences from sequencing artifacts alleviating this bias. 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 assessment study employing a novel dataset and framework indicate that bioinformatic pipeline 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.

The assessment framework consistent of three components, (1) beta-diversity repeatability, (2) biological signal detection, (3) signal detection power. Our assessment framework utilized a novel two-sample titration dataset with multiple levels of technical replication, 16S rRNA PCR, sequencing libraries, and sequencing runs. Multiple PCR replicates were used to assess beta-diversity repeatability, the different sample types (trial participants and exposure status) in conjunction with multiple sequencing runs were used to assess biological signal detection. The titrations were used to assess signal detection power.

Using mean beta diversity between PCR replicates for the four sequencing runs we were able to show that the impact of sequence quality and variation in number of reads on diversity metric repeatability, mean beta diversity between PCR replicates, was pipeline and diversity metric dependent.
For the QIIME De novo pipeline the mean unweighted UniFrac between PCR replicates was high for all runs but low weighted UniFrac. We attributed the difference in results between the weighted and unweighted UniFrac metric results for the QIIME *de novo* dataset to singletons, in ability to group sequencing artifacts with true biological sequences. Singleton removal addresses this bias.
Sequence data from JHU run 1, which had lower error rate and read number variability 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.

To evaluate how well bioinformatic pipeline and normalization methods are able to differentiate between biological signal and noise due to technical variability we compared the mean beta diversity between different biological samples and technical replicates, including PCR and sequencing run. 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 than 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 true biological signal. 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.

For the third component of our assessment we evaluated the relationship between sequence data characteristics, bioinformatic pipeline, and normalization method on sets of samples with varying levels of similarity. The assessment results varied by pipeline and diversity metric, with DADA2 and mothur consistently out performing the other bioinformatic pipelines.
For weighted phylogenetic methods normalization methods rarely improved the results, but improved the results in most cases for weighted taxonomic diversity metrics. Inconsistent results were observed for unweighted metrics, with rarefying data to 5000 total abundance per sample improved the results, rarefying to the 15th percentile lowered performance.

4.1. **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 better able to handle 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 choice of algorithm 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 assessment framework and dataset presented here is novel and can be used to optimize existing pipelines and benchmarking new pipelines.

5. 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.

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.

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. doi: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). doi: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.

Chao, Anne, Nicholas J Gotelli, T C Hsieh, Elizabeth L Sander, K H Ma, Robert K Colwell, and Aaron M Ellison. 2014. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." *Ecol. Monogr.* 84 (1). Ecological Society of America: 45–67.

Chiu, Chun-Huo, and Anne Chao. 2016. "Estimating and Comparing Microbial Diversity in the Presence of Sequencing Errors." *PeerJ* 4 (February): e1634.

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.