Title: To rarefy or not to rarefy: Enhancing microbial community analysis through next-generation sequencing

Authors: Ellen S. Cameron[a], Philip J. Schmidt[b], Benjamin J.-M. Tremblay[a], Monica B. Emelko[b], Kirsten M. Müller[a,*]

[a] Department of Biology, University of Waterloo, 200 University Ave. W, Waterloo, Ontario, Canada, N2L 3G1

[b] Department of Civil & Environmental Engineering, University of Waterloo, 200 University Ave. W, Waterloo, Ontario, Canada, N2L 3G1

[*] Corresponding author. Tel.: +1 519 888 4567x32224

E-mail address: kirsten.muller@uwaterloo.ca (K.M. Müller).

Abstract

The application of amplicon sequencing in water research provides a rapid and sensitive technique for microbial community analysis in a variety of environments ranging from freshwater lakes to water and wastewater treatment plants. It has revolutionized our ability to study DNA collected from environmental samples by eliminating the challenges associated with lab cultivation and taxonomic identification. DNA sequencing data consist of discrete counts of sequence reads, the total number of which is the library size. Samples may have different library sizes and thus, a normalization technique is required to meaningfully compare them. The process of randomly subsampling sequences to a selected normalized library size from the sample library—rarefying—is one such normalization technique. However, rarefying has been criticized as a normalization technique because data can be omitted through the exclusion of either excess sequences or entire samples, depending on the rarefied library size selected. Although it has been suggested that rarefying should be avoided altogether, we propose that repeatedly rarefying enables (i) characterization of the variation introduced to diversity analyses by this random subsampling and (ii) selection of smaller library sizes where necessary to incorporate all samples in the analysis. Rarefying may be a statistically valid normalization technique, but researchers should evaluate their data to make appropriate decisions regarding library size selection and subsampling type. The impact of normalized library size selection and rarefying with or without replacement in diversity analyses were evaluated herein.

Keywords (Maximum 6 keywords)

Library size normalization, amplicon sequencing, alpha diversity, beta diversity, Shannon index, Bray-Curtis dissimilarity

Highlights

- Amplicon sequencing technology for environmental water samples is reviewed
- Sequencing data must be normalized to allow comparison in diversity analyses
- Rarefying normalizes library sizes by subsampling from observed sequences
- Criticisms of data loss through rarefying can be resolved by rarefying repeatedly
- Rarefying repeatedly characterizes errors introduced by subsampling sequences

1. Introduction

Next-generation sequencing has revolutionized the analysis of environmental systems through the characterization of microbial communities and their function by the study of DNA collected from samples that contain mixed assemblages of organisms (Bartram et al., 2011; Hugerth and Andersson, 2017; Shokralla et al., 2012). Fewer than 1% of species in the environment can be isolated and cultured, limiting the ability to identify rare and difficult-to-cultivate members of the community (Bodor et al., 2020; Cho and Giovannoni, 2004; Ferguson et al., 1984). In addition to the limitations of culturing, microscopic evaluation of environmental samples remains of limited utility because of challenges in high-resolution taxonomic identification and the inability to infer function from morphology (Hugerth and Andersson, 2017). Metagenomic studies employ next-generation sequencing technology to analyze environmental DNA (Thomas et al., 2012) and have largely eliminated these challenges (McMurdie and Holmes, 2014).

Metagenomics includes a conglomerate of different sequencing experimental designs including amplicon sequencing (sequencing of amplified genes of interest) and shotgun sequencing (the sequencing of fragments of present genetic material). While shotgun sequencing allows characterization of the entire community, including both taxonomic composition and functional gene profiles, it is not widely accessible due to high sequencing costs and computational requirements for analysis (Bartram et al., 2011; Clooney et al., 2016; Langille et

al., 2013). In contrast, the relatively low cost of amplicon sequencing has made it an increasingly popular technique (Clooney et al., 2016; Langille et al., 2013). The amplification and sequencing of specific genes (i.e., taxonomic marker genes) enables characterization of microbial community composition (Hodkinson and Grice, 2015); as a result, it has been successfully applied in many area of water research. For example, amplicon sequencing has been used to characterize and predict cyanobacteria blooms (Tromas et al., 2017), describe microbial communities found in aquatic ecosystems (Zhang et al., 2020), and evaluate groundwater vulnerability to pathogen intrusion (Chik et al., 2020). It has also been applied to water quality and treatment performance monitoring in diverse settings (Vierheilig et al., 2015) including drinking water distribution systems (Perrin et al., 2019; Shaw et al., 2015), drinking water biofilters (Kirisits et al., 2019), anaerobic digesters (Lam et al., 2020), and cooling towers (Paranjape et al., 2020).

Processing and analysis of amplicon sequencing data is statistically complicated for a number of reasons (Weiss et al., 2017). For example, library sizes (i.e., the total number of sequencing reads within a sample) can vary widely among different samples, even within a single sequencing run, and the disparity in library sizes between samples may not represent actual differences in microbial communities (McMurdie and Holmes, 2014). Further, sequence counts cannot be compared directly among samples with variable library sizes (Gloor et al., 2016). For example, a sample with 5,000 sequence reads is likely to have a different read count for a specific sequence variant than a sample with 20,000 sequence reads simply due to the differences in library size. In order to draw biologically meaningful conclusions from amplicon sequencing data, samples must be normalized to remove the artificial detection of variation between samples

due to differences in library sizes. Notably, a variety of normalization techniques that may affect the analysis and interpretation of results have been suggested.

Rarefaction is a normalization tool initially developed for ecological diversity analyses to allow for sample comparison without associated bias from differences in sample size (Sanders, 1968). Rarefaction normalizes samples of differing sample size by subsampling to a normalized threshold, often equal to the smallest sample size (Sanders, 1968; Willis, 2019). For samples that are larger than the normalized threshold, data is discarded until the normalized threshold value is met (Willis, 2019). Although initially developed for use in ecological studies, rarefaction is a commonly used and widely debated library size normalization technique for amplicon sequencing data (Gloor et al., 2017; McMurdie and Holmes, 2014). Similarly to the original employment in ecological studies, libraries are subsampled to create “rarefied” libraries of a consistent size among samples (Gloor et al., 2017; McMurdie and Holmes, 2014). The application of rarefying as a library size normalization technique has been critiqued due to the artificial variation introduced through subsampling and the omission of valid data through loss of sequence counts or exclusion of samples with small library sizes (McMurdie and Holmes, 2014). Typically, rarefying of sequences is only conducted once to provide a snapshot of the community at the smaller, normalized library size; thus, this approach does not assess variability introduced through subsampling. Due to the concerns associated with random subsampling that results in loss of data and generation of artificial variation, other normalization approaches featuring various statistical transformations have been proposed (Badri et al., 2018; Chen et al., 2018; Gloor et al., 2017; McMurdie and Holmes, 2014). Examples include expressing counts as proportions of the total library size (McMurdie and Holmes, 2014), centered log-ratio

transformations (Gloor et al., 2017), geometric mean pairwise ratios (Chen et al., 2018), variance stabilizing transformations and relative log expressions (Badri et al., 2018).

Some of the proposed alternatives to rarefying are compromised by the presence of zeros, often for large numbers of variants of interest. For example, the centered log-ratio transformation cannot be applied to datasets with zero counts (Gloor et al., 2017). Using such transformations may involve the omission of zero count sequence variants from libraries, substitution of a pseudocount in place of zeros, or modelling zero counts through probabilistic distributions (Gloor et al., 2017, 2016). Selecting an appropriate pseudocount value is an additional challenge (Weiss et al., 2017). Zero counts represent a lack of information (Silverman et al., 2018) and may arise from true absence of the sequence variant in the sample, or a loss resulting in it not being detected when it was actually present (Tsilimigras and Fodor, 2016; Wang and LêCao, 2019). Zeros are a natural occurrence in discrete count-based data such as the counting of microorganisms or amplicon sequences, and adjusting them can introduce substantial bias into microbial analyses (Chik et al., 2018). While the 16S rRNA gene is the standard in taxonomic marker gene analysis of organisms within the domains Bacteria and Archaea, there is a need to identify and develop consistent approaches for statistical analysis and interpretation of this amplicon data.

Here, the application of rarefying as a library size normalization technique is investigated to determine if criticisms of the technique remain when rarefying is implemented repeatedly, which allows evaluation of the variability introduced in subsampling from the original library size to a lower rarefied library size shared among all samples. Specifically, this paper addresses (i) appropriate usage of rarefying to characterize variation introduced through random subsampling, (ii) the impact of subsampling with or without replacement on diversity analysis,

and (iii) the impact of library size selection on diversity analyses such as the Shannon index and Bray-Curtis dissimilarity ordinations. Different scenarios were generated to evaluate the impacts of rarefying library sizes on the interpretation of community diversity analyses.

2. Background

Due to the inevitable interdisciplinarity of water research and the complexity and novelty of next generation sequencing relative to traditional microbiological methods used in water quality analyses, further detail on amplicon sequencing is provided. Amplification and sequencing of taxonomic marker genes, such as the 16S rRNA gene in mitochondria, chloroplasts, bacteria and archaea (Case et al., 2007; Tsukuda et al., 2017; Weisburg et al., 1991; Yang et al., 2016) and the 18S rRNA gene within the nucleus of eukaryotes (Field et al., 1988), has been used extensively to examine phylogeny, evolution and taxonomic classification of numerous groups across the three domains of life (Quast et al., 2013; Weisburg et al., 1991; Woese et al., 1990). Widely-used reference databases have been developed containing marker gene sequences across numerous phyla (Hugerth and Andersson, 2017).

The 16S rRNA gene consists of nine highly conserved regions separated by nine hypervariable regions (V1-V9) (Gray et al., 1984) and is approximately 1,540 base pairs in length (Kim et al., 2011; Schloss and Handelsman, 2004). While sequencing of the full 16S rRNA gene provides the highest taxonomic resolution (Johnson et al., 2019), many studies only utilize partial sequences of the 16S rRNA gene due to limitations in read length in NGS platforms (Kim et al., 2011). NGS on Illumina platforms produces paired end reads up to 350 base pairs in length, requiring researchers to select an appropriate region of the 16S rRNA gene to amplify and sequence to optimize taxonomic resolution with partial 16S rRNA gene sequences (Bukin et al., 2019; Kim et al., 2011). Sequencing the more conservative regions of

the 16S rRNA gene may provide better resolution for higher levels of taxonomy, while more variable regions can provide higher resolution for the classification of sequences to the genus and species levels in bacteria and archaea (Bukin et al., 2019; Kim et al., 2011; Yang et al., 2016). Different variable regions of the 16S rRNA gene may be biased towards different taxa (Johnson et al., 2019) and be preferred for different ecosystems (Escapa et al., 2020). For example, the V4 region has been shown to strongly differentiate taxa from the phyla Cyanobacteria, Firmicutes, Fusobacteria, Plantomycetes, and Tenericutes but the V3 region best differentiates taxa from the phyla Proteobacteria (e.g., *Escherichia coli*, *Salmonella*, *Campylobacter*), Acidobacteria, Bacteroidetes, Chloroflexi, Gemmatimonadetes, Nitrospirae, and Spirochaetae and Proteobacteria (Zhang et al., 2018). The V4 region of the 16S rRNA gene is frequently targeted using specific primers designed to minimize phylum amplification bias accounting for common aquatic bacteria (Walters et al., 2015) and is frequently used in aquatic studies (Zhang et al., 2018). It is important to consider suitability of a 16S rRNA region for the habitat (Escapa et al., 2020) and the taxa present in the microbial community due to potential bias of analyzing differing subregions of the 16S rRNA gene (Johnson et al., 2019; Zhang et al., 2018).

The use of amplicon sequencing of partial sequences of the 16S rRNA gene allows examination of microbial community composition and the exploration of shifts in community structure, correlations between taxa and environmental factors (Hodkinson and Grice, 2015), and identification of differentially abundant taxa between samples (Hugerth and Andersson, 2017). Amplicon sequencing datasets can be analyzed using a variety of bioinformatic platforms for sequence analysis (e.g., sequence denoising, taxonomic classification, diversity analysis) including *mothur* (Schloss et al., 2009) and *QIIME2* (Bolyen et al., 2019). Previously, sequencing analysis involved the creation of operational taxonomic units (OTUs) by clustering

sequences into groups that met a certain similarity threshold, resulting in a loss of representation of variation in sequences (Callahan et al., 2017). Additionally, the generation of OTUs are dependent on the dataset from which they were generated, precluding cross-study comparison (Callahan et al., 2017). Advances in computational power have allowed a shift from use of OTUs to amplicon sequence variants (ASVs) representative of each unique sequence in a sample (Callahan et al., 2017). The use of ASVs allows comparisons of sequence variants generated in different studies and does not lose the representation of biological variation within OTUs (Callahan et al., 2017). The implementation of tools included in bioinformatics pipelines, such as DADA2 (Callahan et al., 2016) or Deblur (Amir et al., 2017), allows quality control of sequencing through the removal of sequencing errors and for the production of ASVs (Amir et al., 2017; Callahan et al., 2016). Taxonomic classification of 16S rRNA sequences using rRNA databases including SILVA (Quast et al., 2013), the Ribosomal Database Project (Cole et al., 2014) and GreenGenes (DeSantis et al., 2006) allows researchers to construct taxonomic community profiles (Bartram et al., 2011). Quality controlled sequencing data for a particular run is then organized into large matrices where columns represent experimental samples and rows contain counts for different ASVs (Weiss et al., 2017). Amplicon sequencing samples will have a total number of sequencing reads, also known as the library size (McMurdie and Holmes, 2014) but does not provide information on the absolute abundance of sequence variants (Gloor et al., 2017, 2016).

Amplicon sequencing data can be used for studies on taxonomic composition, differential abundance analysis and diversity analyses (Figure 1). Exploring the taxonomic composition allows characterization of microbial communities according to the classification of sequence variants based on similarities to sequences in online databases. The use of taxonomic

composition with sequencing data is regularly expressed in proportions or relative abundances to allow comparison of the community composition without the bias introduced from having differing library sizes between samples. Differential abundance analysis may be utilized to explore whether specific sequence variants are found in significantly different abundances between samples (Weiss et al., 2017) to identify potential biological drivers for these differences in read counts but will not be explored in this article. Differential abundance analysis is frequently performed using programs initially designed for transcriptomics such as DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2009) or programs designed to account for the compositional structure of sequence data ALDeX2 (Fernandes et al., 2014). The true diversity of an environmental sample is unknown making diversity analyses on microbial communities inherently difficult due to the inability to provide an absolute enumeration that accurately reflects the microbial community (Hughes et al., 2001).

Without knowing the true diversity of microbial samples, accumulation or rarefaction curves are frequently utilized to examine the observed sequence variants in relation to the library size (Hughes et al., 2001). The rarefaction curve will inform researchers on general trends on the maximal diversity observed with the expectation that as curves approach an asymptote, the full diversity of the community has been captured in sequencing efforts (Hughes et al., 2001). Diversity analyses are broadly categorized as alpha diversity (within sample diversity), beta diversity (between samples) or gamma-diversity (between samples on a large spatial scale) (Sepkoski, 1988). Alpha diversity and beta diversity are largely employed in the studies of microbial communities in amplicon sequencing datasets to explore the diversity found in samples and contrast the community composition observed between samples.

Alpha diversity serves to identify richness (e.g., number of observed sequence variants) and evenness (e.g., allocation of read counts across observed sequence variants) within a sample (Willis, 2019). Comparison of alpha diversity with samples of differing library sizes may result in inherent biases with samples having larger library sizes appearing more diverse than samples with smaller library sizes due to the presence of more potential sequence variants in samples with larger libraries (Willis, 2019). This requires samples to have equivalent library sizes before comparison to prevent the risk of observing inherent bias fabricated only from (Willis, 2019). Diversity indices used to characterize the alpha diversity of samples include the Shannon index (Shannon, 1948), Chao1 index (Chao and Bunge, 2002), and abundance coverage estimator (Hughes et al., 2001), but unique details of such indices should be understood for correct usage. For example, Chao1 relies on the observation of singletons in data to estimate diversity (Chao and Bunge, 2002), but denoising processes for sequencing data may remove singleton reads making the Chao1 estimator invalid for accurate analysis.

Similar to alpha diversity, samples with differing library sizes in beta diversity analyses may produce erroneous results due to the potential for samples with larger library sizes to be observed as having more unique sequences simply due to the presence of more sequence variants (Weiss et al., 2017). A variety of different beta diversity metrics can be used to compare sequence variant composition between samples including Bray-Curtis (Bray and Curtis, 1957) or Unifrac (Lozupone and Knight, 2007) distances and then visualized using ordination techniques.

3. Methods

3.1 Sample Collection, DNA Extraction & Amplicon Sequencing

Samples used in the analyses are part of a larger study at Turkey Lakes Watershed (North Part, ON) but only an illustrative subset of samples is considered for the purpose of evaluating

rarefying as a normalization tool herein. The water samples (1 L) were collected from oligotrophic lakes located in the Canadian Shield in Ontario, Canada and were vacuum filtered through 47 mm Whatman GF/C filters (Whatman plc, Buckinghamshire, United Kingdom). DNA was extracted from ground filter material using the DNeasy PowerSoil Kit (QIAGEN Inc., Venlo, Netherlands) following the kit protocols. DNA extracts were submitted for amplicon sequencing using the Illumina MiSeq platform (Illumina Inc., San Diego, United States) at the commercial laboratory Metagenom Bio Inc. (Waterloo, ON). Primers designed to target the 16S rRNA gene V4 region [515FB (GTGYCAGCMGCCGCGGTAA) and 806RB (GGACTACNVGGGTWTCTAAT)] (Walters et al., 2015) were used for PCR amplification.

3.2 Sequence Processing & Library Size Normalization

The program QIIME2 (v. 2019.10; Bolyen et al., 2019) was used for bioinformatic processing of sequence reads. Demultiplexed paired-end sequences were trimmed and denoised, including the removal of chimeric sequences and singleton sequence variants to remove sequences that may not be representative of real organisms, using DADA2 (Callahan et al., 2016) to construct the ASV table. Taxonomic classification was performed using a classifier trained with the SILVA138 database (Quast et al., 2013; Yilmaz et al., 2014). Output files from QIIME2 were imported into R (v. 4.0.1) (R Core Team, 2020) for community analyses using *qiime2R* (v. 0.99.23) (Bisanz, 2018). Initial sequence libraries were further filtered using *phyloseq* (v. 1.32.0) (McMurdie and Holmes, 2013) to exclude amplicon sequence variants that were taxonomically classified as mitochondria or chloroplast sequences. We developed a package called *mirlyn* (Multiple Iterations of Rarefaction for Library Normalization; Cameron and Tremblay, 2020) that facilitates implementation of techniques used in this study built from existing R packages (Table S1). Using the output from *phyloseq*, *mirlyn* was used to generate rarefaction curves, to

repeatedly rarefy libraries to account for variation in library sizes among samples, and to plot diversity metrics given repeated rarefaction.

3.3 Community Diversity Analyses on Normalized Libraries

The Shannon index, an alpha diversity metric, was analyzed for sample comparison as a function of normalized library size. Normalized libraries were also used for beta diversity analysis. A Hellinger transformation was applied to normalized libraries to account for the arch effect regularly observed in ecological count data and Hellinger transformed data were then used to calculate Bray-Curtis distances (Bray and Curtis, 1957). Principal component analysis (PCA) was conducted on the Bray-Curtis distance matrices.

3.4 Study Approach

Typically, rarefaction has only been conducted a single time in microbial community analyses, and this omits a random subset of observed sequences introducing a possible source of error. To examine this error, samples were repeatedly rarefied 1000 times. This repetition provided a representative suite of rarefied samples capturing the variation in sequence variant composition imposed by rarefying. The sections below address the various decisions that must be made by the analyst and factors affecting reliability of results when rarefaction is used.

3.4.1 The Effects of Subsampling Approach – With or Without Replacement

Rarefying library sizes may be performed with or without replacement. To evaluate the effects of subsampling replacement approaches, we repeatedly rarefied filtered sequence libraries to varying depths with and without replacement. Results of the two approaches were contrasted in diversity analyses to evaluate the impact of subsampling approach on interpretation of results.

3.4.2 The Effects of Normalized Library Size Selection

Rarefying involves the selection of an appropriate sampling depth to be shared by each sample. To evaluate the effects of different rarefied library sizes, filtered sequence libraries were rarefied repeatedly without replacement to varying depths. Results for varying sampling depths were contrasted in diversity analyses to evaluate the impact of normalized library size selection on interpretation of results.

4. Results & Discussion

4.1 Use of Rarefaction Curves to Explore Suitable Normalized Library Sizes

Rarefying requires the selection of a potentially arbitrary normalized library size, which can impact subsequent community diversity analyses and therefore presents users with the challenge of making an appropriate decision of what size to select (McMurdie and Holmes, 2014). Suitable sampling depths for groups of samples can be determined through the examination of rarefaction curves (Figure 2). By selecting a library size that encompasses the flattening portion of the curve for each sample, it is generally assumed that the normalized library size will adequately capture the diversity within the samples despite the exclusion of sequence variants during the rarefying process (i.e., there are progressively diminishing returns in including more of the observed sequence variants as the rarefaction curve flattens).

Suggestions have previously been made encouraging selection of a normalized library size that is encompassing of most samples (e.g., 10,000 sequences) and advocation against rarefying below certain depths (e.g., 1,000 sequences) due to decreases in data quality (Navas-Molina et al., 2013). However, generic criteria may not be applicable to all datasets and exploratory data analysis is often required to make informed and appropriate decisions on the selection of a normalized library size. Although previous research advises against rarefying

below certain thresholds, users may be presented with the dilemma of selecting a sampling depth that either does not capture the full diversity of a sample depicted in the rarefaction curve (Figure 2 – I) or would require the omission of entire samples with smaller library sizes (Figure 2 – III). The implementation of multiple iterations of rarefying library sizes will aid in alleviating this dilemma by capturing the potential losses in community diversity for samples that are rarefied to lower than ideal depth.

4.2 The Effects of Subsampling Approach & Normalized Library Size Selection on Alpha Diversity Analyses

The differences in input parameters for rarefying samples requires users to be diligent in the selection of appropriate tools and commands for their analysis. The R package *phyloseq*, a popular tool for microbiome analyses, has default settings for rarefying including sampling with replacement to optimize computational run time and memory usage (McMurdie and Holmes, 2013), although sampling without replacement is more appropriate statistically. Sampling without replacement draws a subset from the observed set of sequences, whereas sampling with replacement fabricates a set of sequences in similar proportions to the observed set of sequences (Figure 3). Sampling with replacement can potentially cause a rare sequence variant to appear more frequently in the rarefied sample than it was actually observed in the original library.

Rarefying libraries with or without replacement was not found to substantially impact the Shannon index in the scenarios considered in this study (Figure 4-A), but users should still be aware of potential implications of sampling with or without replacement when rarefying libraries. Libraries rarefied with replacement are observed to have a slightly reduced Shannon index relative to libraries rarefied without replacement at many library sizes.

The conservation of larger library sizes when rarefying allows detection of more diversity with minimal variation observed between the iterations of rarefaction (Figure 4-A). The largest considered library size (11,213 sequences) captured the highest Shannon index value, while the Shannon index diminishes at lower normalized library sizes. The use of multiple repeated iterations of rarefying allows variation introduced through subsampling to be represented in the diversity metric and the selection of larger library sizes maintains precision in the calculation of the diversity metric. While there was only slight disparity in the Shannon index values between the largest library size and unnormalized data, this may not always be the case and is dependent on the sequence variant composition of the samples. Samples dominated by a large number of low abundance sequence variants are more likely to have a reduced Shannon index value at a larger normalized library size. Alternatively, samples dominated by only several highly abundant sequence variants will be comparatively robust to rarefying. A plot of the Shannon index as a function of rarefied library size (Figure 4-B) demonstrates the overall robustness of the Shannon index of these samples for larger library sizes (e.g., > 5,000 sequences) and the increased variation and diminishing values when proceeding to smaller rarefied library sizes. When the normalized library size was decreased to 5,000, the Shannon index is still only slightly reduced by the rarefaction but there is a decrease in precision of the value due to introduced variation from rarefying.

The precision of the diversity metric is extremely degraded when libraries were rarefied to the smallest considered library size of 500 sequences. The degradation of precision observed in the diversity metric may result in the inappropriate claim of identical diversity values between samples that were shown to have varying diversity values at larger normalized library sizes. The extreme loss of precision of the Shannon index suggests that the selection of smaller library sizes

should be approached with caution when using alpha diversity metrics while larger normalized library sizes prevent loss of precision and reduction of the Shannon index value. However, as previously noted, the reduction in the value of the Shannon index will be dependent on the sequence variant composition of the samples. When rarefying is implemented as a library size normalization technique for alpha diversity analyses, it is informative to perform multiple repeated iterations to capture the artificial variation introduced into libraries through random subsampling.

Previous research conducted evaluating library size normalization techniques have only focused on beta diversity analysis and differential abundance analysis (Gloor et al., 2017; McMurdie and Holmes, 2014; Weiss et al., 2017) but the appropriateness of library size normalization techniques for alpha diversity metrics must be evaluated due to the prerequisite of having equal library sizes for accurate calculation. Utilization of unnormalized library sizes with alpha diversity metrics may generate bias due to the potential for samples with larger library sizes to be inherently more diverse than a sample with a small library size. The repeated iterations of rarefying library sizes allows characterization of uncertainty in sample diversity at any rarefied library size (Figure 3).

4.3 The Effects of Subsampling Approach & Normalized Library Size Selection on Beta Diversity Analysis

When samples were repeatedly rarefied to a common normalized library size with and without replacement (Figure 5), similar amounts of variation in the Bray-Curtis PCA ordinations were observed between the sampling approaches. This indicates that although rarefying with replacement seems potentially erroneous due to the fabrication of count values that are not representative of actual data, the impact on the variation introduced into the Bray-Curtis

dissimilarity distances is not large and will likely not interfere with the interpretation of results. However, rarefying without replacement should be encouraged because it is more theoretically correct, and it has not been comprehensively demonstrated that sampling with replacement is a valid approximation for all types of diversity analysis or library compositions.

When larger normalized library sizes are maintained through rarefaction, there is less potential variation introduced into beta diversity analyses, including Bray-Curtis dissimilarity PCA ordinations (Figure 5). For example, in the largest normalized library size (Figure 5A) possible for these data, minimal amounts of variation were observed within each community, indicating that the preservation of higher sequence counts minimizes the amount of artificial variation introduced into datasets by rarefaction (including no variation for Sample F because it is not actually rarefied in this scenario). For this reason, rarefying to the smallest library size of a set of samples is a sensible guideline. Although, a normalized library size of 5,000 is lower than the flattening portion of the rarefaction curve for samples A, B, and C (Figure 2), the selection of this potentially inappropriate normalized library size (Figure 5C) can still accurately reflect the diversity between samples without excess artificial variation introduced through rarefaction. Due to the variation in the Bray-Curtis dissimilarity ordinations in the smaller library sizes (Figure 5E/G), it is critical to include computational replicates of rarefied libraries to fully characterize the introduced variation in communities. Without repeated replication, rarefaction to small normalized library sizes could result in artificial similarity or dissimilarity identified between samples.

Beta diversity analyses of very small rarefied libraries without replacement (Figure 6A, B, C) can still reflect similar clustering patterns observed in larger library sizes but with a much lower resolution of clusters. Rarefying has previously been shown to be an appropriate

normalization tool for samples with low sequence counts (e.g., <1,000 sequences per sample) (Weiss et al., 2017), which is promising for datasets containing samples with small initial library sizes. However, selection of a smaller-than-necessary normalized library size will introduce excess artificial variation into beta diversity analyses and will needlessly interfere with interpretation of results. Caution must be taken to avoid selection of an excessively small normalized library size due to the introduction of extreme levels of artificial variation that compromises accurate depiction of diversity (Figure 6D) and only reflects small portions of the sequence variants from samples with large library sizes. The tradeoff between rarefying to a smaller than advisable library size or excluding entire low sequence count samples remains and can possibly be resolved by analyzing results with all samples and a small rarefied library size as well as with some omitted samples and a larger rarefied library size.

Although rarefying has the potential to introduce artificial variation into data used in beta diversity analyses, these results suggest that it does not become problematic until rarefying to normalized library sizes that are very small (e.g., 500 sequences or less). While in alpha diversity analyses we saw a disintegration of the precision and resolution of the Shannon index at 500 sequences, beta diversity analyses may be more robust to rarefaction and capable of reflecting qualitative clusters in ordination as previously discussed in Weiss et al. (2017). The artificial variation introduced to beta diversity analyses by rarefaction could lead to erroneous interpretation of results, but the implementation of multiple iterations of rarefying library sizes allows a full representation of this variation, further supporting rarefying library sizes as a statistically valid tool despite criticisms.

One of the largest criticisms of rarefying as a normalization technique is the omission of valid data with the suggestion that repeated iterations still does not provide stability (McMurdie

and Holmes, 2014); however, the repeated iterations of rarefying in this study demonstrated that rarefying repeatedly does not substantially impact the output and interpretation of beta diversity analyses unless rarefying to sizes that are inadvisable to begin with. Furthermore, supporting material of McMurdie and Holmes (2014) evaluating the use of repeated rarefying (100 replicates) critiqued the implementation of repeatedly rarefying due to the generation of artificial variation; however, the repeated iterations allow characterization of uncertainty while accounting for discrepancies in library size. The additional criticism of rarefying due to the generation of artificial variation (McMurdie and Holmes, 2014) has been demonstrated in this study to only be a concern for beta diversity analyses when rarefying to very small sizes, but the inclusion of computational repetitions in analyses provides a full representation of any possible artificial variation introduced through subsampling. The use of non-normalized data or proportional data has been shown to be more susceptible to the generation of artificial clusters in ordinations and rarefying has been demonstrated to be an effective normalization technique for beta diversity analyses (Weiss et al., 2017). The efficacy of rarefying as a normalization tool for beta diversity analysis demonstrated herein further supports the use of rarefied library sizes with multiple iterations to characterize artificial variation unavoidably introduced to the rarefied community composition.

Conclusions

- Rarefying with or without replacement did not substantially impact the interpretation of alpha (Shannon index) or beta (Bray-Curtis dissimilarity) diversity analyses considered in this study but rarefying without replacement will provide more accurate reflection of sample diversity.

- To avoid the arbitrary loss of available information through rarefaction to a common library size, the random error introduced through rarefaction may be evaluated by repeating rarefaction multiple times.
- The use of larger normalized library sizes when rarefying minimizes the amount of artificial variation introduced into diversity analyses but may necessitate omission of samples with small library sizes (or analysis at both inclusive low library sizes and restrictive higher library sizes).
- Ordination patterns are relatively well preserved down to small normalized library sizes with increasing variation shown by repeatedly rarefying, whereas the Shannon index is very susceptible to being impacted by small normalized library sizes both in declining values and variability introduced through rarefaction.
- Even though repeated rarefaction can characterize the error introduced by excluding some fraction of the sequence variants, rarefying to extremely small sizes (e.g., 100 sequences) is inappropriate due to large error and inability to differentiate between sample clusters.

Acknowledgements

We acknowledge the support of the forWater NSERC network for forested drinking water source protection technologies [NETGP-494312-16]. We are also grateful for the continued support of Natural Resources Canada and Environment and Climate Change Canada in sample collection at Turkey Lakes Watershed Research Station

4. References

Amir, A., Daniel, M., Navas-Molina, J., Kopylova, E., Morton, J., Xu, Z.Z., Eric, K., Thompson, L., Hyde, E., Gonzalez, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191, e00191-16. https://doi.org/https://doi.org/10.1128/mSystems.00191-16

Badri, M., Kurtz, Z., Muller, C., Bonneau, R., 2018. Normalization methods for microbial abundance data strongly affect correlation estimates. bioRxiv 406264. https://doi.org/10.1101/406264

Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., Neufeld, J.D., 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. 77, 3846–3852. https://doi.org/10.1128/AEM.02772-10

Bisanz, J.E., 2018. qiime2R: Importing QIIME2 artifacts and associated data into R sessions. https://github.com/jbisanz/qiime2R.

Bodor, A., Bounedjoum, N., Vincze, G.E., Erdeiné Kis, Á., Laczi, K., Bende, G., Szilágyi, Á., Kovács, T., Perei, K., Rákhely, G., 2020. Challenges of unculturable bacteria: environmental perspectives. Rev. Environ. Sci. Biotechnol. 19, 1–22. https://doi.org/10.1007/s11157-020-09522-4

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K. Bin, Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A. V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bray, J.R., Curtis, J.T., 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. 27, 325–349. https://doi.org/10.2307/1942268

Bukin, Y.S., Galachyants, Y.P., Morozov, I. V., Bukin, S. V., Zakharenko, A.S., Zemskaya, T.I.,

2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci. Data 6, 1–14. https://doi.org/10.1038/sdata.2019.7

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Cameron, E.S., Tremblay, B.J.-M. 2020. mirlyn: Multiple Iterations of Rarefying for Library Normalization. https://github.com/escamero/mirlyn/

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., Kjelleberg, S., 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl. Environ. Microbiol. 73, 278–288. https://doi.org/10.1128/AEM.01177-06

Chao, A., Bunge, J., 2002. Estimating the number of species in a stochastic abundance model. Biometrics 58, 531–539. https://doi.org/10.1111/j.0006-341X.2002.00531.x

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., Chen, J., 2018. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ 2018, 1–20. https://doi.org/10.7717/peerj.4600

Chik, A.H.S., Emelko, M.B., Anderson, W.B., O'Sullivan, K.E., Savio, D., Farnleitner, A.H., Blaschke, A.P., Schijven, J.F., 2020. Evaluation of groundwater bacterial community composition to inform waterborne pathogen vulnerability assessments. Sci. Total Environ. 743, 140472. https://doi.org/10.1016/j.scitotenv.2020.140472

Chik, A.H.S., Schmidt, P.J., Emelko, M.B., 2018. Learning something from nothing: The critical importance of rethinking microbial non-detects. Front. Microbiol. 9, 1–9. https://doi.org/10.3389/fmicb.2018.02304

Cho, J.C., Giovannoni, S.J., 2004. Cultivation and Growth Characteristics of a Diverse Group of Oligotrophic Marine Gammaproteobacteria. Appl. Environ. Microbiol. 70, 432–440. https://doi.org/10.1128/AEM.70.1.432-440.2004

Clooney, A.G., Fouhy, F., Sleator, R.D., O'Driscoll, A., Stanton, C., Cotter, P.D., Claesson, M.J., 2016. Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. PLoS One 11, 1–16. https://doi.org/10.1371/journal.pone.0148028

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M., 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, 633–642. https://doi.org/10.1093/nar/gkt1244

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072. https://doi.org/10.1128/AEM.03006-05

F Escapa, I., Huang, Y., Chen, T., Lin, M., Kokaras, A., Dewhirst, F.E., Lemon, K.P., 2020. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8, 65. https://doi.org/10.1186/s40168-020-00841-w

Ferguson, R.L., Buckley, E.N., Palumbo, A. V., 1984. Response of marine bacterioplankton to differential filtration and confinement. Appl. Environ. Microbiol. 47, 49–55. https://doi.org/10.1128/aem.47.1.49-55.1984

Fernandes, A.D., Reid, J.N.S., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B., 2014. Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2, 1–13. https://doi.org/10.1186/2049-2618-2-15

Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., Raff, R.A., 1988. Molecular phylogeny of the animal kingdom. Science (80-.). 239, 748–753. https://doi.org/10.1126/science.3277277

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome datasets are compositional: And this is not optional. Front. Microbiol. 8, 1–6. https://doi.org/10.3389/fmicb.2017.02224

Gloor, G.B., Macklaim, J.M., Vu, M., Fernandes, A.D., 2016. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian J. Stat. 45, 73–87. https://doi.org/10.17713/ajs.v45i4.122

Gray, M.W., Sankoff, D., Cedergren, R.J., 1984. On the evolutionary descent of organisms and organelles: A global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. Nucleic Acids Res. 12, 5837–5852. https://doi.org/10.1093/nar/12.14.5837

Hall, M., Beiko, R.G., 2018. 16S rRNA Gene Analysis with QIIME2, in: Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-8728-3_8

Hodkinson, B.P., Grice, E.A., 2015. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv. Wound Care 4, 50–58. https://doi.org/10.1089/wound.2014.0542

Hugerth, L.W., Andersson, A.F., 2017. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. Front. Microbiol. 8, 1561. https://doi.org/10.3389/fmicb.2017.01561

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., Bohannan, B.J.M., 2001. Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. Appl. Environ. Microbiol. 67, 4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001

Johnson, J.S., Spakowicz, D.J., Hong, B.Y., Petersen, L.M., Demkowicz, P., Chen, L., Leopold, S.R., Hanson, B.M., Agresta, H.O., Gerstein, M., Sodergren, E., Weinstock, G.M., 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat. Commun. 10, 1–11. https://doi.org/10.1038/s41467-019-13036-1

Kim, M., Morrison, M., Yu, Z., 2011. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. J. Microbiol. Methods 84, 81–87. https://doi.org/10.1016/j.mimet.2010.10.020

Kirisits, M.J., Emelko, M.B., Pinto, A.J., 2019. Applying biotechnology for drinking water biofiltration: advancing science and practice. Curr. Opin. Biotechnol. 57, 197–204. https://doi.org/10.1016/j.copbio.2019.05.009

Lam, T.Y.C., Mei, R., Wu, Z., Lee, P.K.H., Liu, W.T., Lee, P.H., 2020. Superior resolution characterisation of microbial diversity in anaerobic digesters using full-length 16S rRNA gene amplicon sequencing. Water Res. 178, 115815. https://doi.org/10.1016/j.watres.2020.115815

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower, C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31, 814–821. https://doi.org/10.1038/nbt.2676

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U. S. A. 104, 11436–11440. https://doi.org/10.1073/pnas.0611525104

McMurdie, P.J., Holmes, S., 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput. Biol. 10. https://doi.org/10.1371/journal.pcbi.1003531

McMurdie, P.J., Holmes, S., 2013. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 8, e61217. https://doi.org/10.1371/journal.pone.0061217

Navas-Molina, J.A., Peralta-Sánchez, J.M., González, A., McMurdie, P.J., Vázquez-Baeza, Y., Xu, Z., Ursell, L.K., Lauber, C., Zhou, H., Song, S.J., Huntley, J., Ackermann, G.L., Berg-Lyons, D., Holmes, S., Caporaso, J.G., Knight, R., 2013. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol. 531, 371–444. https://doi.org/10.1016/B978-0-12-407863-5.00019-8

Paranjape, K., Bédard, É., Whyte, L.G., Ronholm, J., Prévost, M., Faucher, S.P., 2020. Presence of Legionella spp. in cooling towers: the role of microbial diversity, Pseudomonas, and continuous chlorine application. Water Res. 169, 115252. https://doi.org/10.1016/j.watres.2019.115252

Perrin, Y., Bouchon, D., Delafont, V., Moulin, L., Héchard, Y., 2019. Microbiome of drinking water: A full-scale spatio-temporal study to monitor water quality in the Paris distribution system. Water Res. 149, 375–385. https://doi.org/10.1016/j.watres.2018.11.013

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, 590–596. https://doi.org/10.1093/nar/gks1219

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

https://doi.org/10.1093/bioinformatics/btp616

Sanders, H.L., 1968. Marine Benthic Diversity : A Comparative Study. Am. Nat. 102, 243–282.

Schloss, P.D., Handelsman, J., 2004. Status of the Microbial Census. Microbiol. Mol. Biol. Rev. 64, 686–691. https://doi.org/10.1128/mmbr.68.4.686-691.2004

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Sepkoski, J.J., 1988. Alpha, beta, or gamma: Where does all the diversity go? Paleobiology 14, 221–234. https://doi.org/10.1017/S0094837300011969

Shaw, J.L.A., Monis, P., Weyrich, L.S., Sawade, E., Drikas, M., Cooper, A.J., 2015. Using amplicon sequencing to characterize and monitor bacterial diversity in drinking water distribution systems. Appl. Environ. Microbiol. 81, 6463–6473. https://doi.org/10.1128/AEM.01297-15

Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 21, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Silverman, J., Roche, K., Mukherjee, S., David, L., 2018. Naught all zeros in sequence count data are the same. bioRxiv 477794. https://doi.org/10.1101/477794

Thomas, T., Gilbert, J., Meyer, F., 2012. Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. 2, 3. https://doi.org/10.1186/2042-5783-2-3

Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., Greer, C.W., Shapiro, B.J., 2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. ISME J. 11, 1746–1763. https://doi.org/10.1038/ismej.2017.58

Tsilimigras, M.C.B., Fodor, A.A., 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann. Epidemiol. 26, 330–335. https://doi.org/10.1016/j.annepidem.2016.03.002

Tsukuda, M., Kitahara, K., Miyazaki, K., 2017. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs. Sci. Rep. 7, 1–8. https://doi.org/10.1038/s41598-017-10214-3

Vierheilig, J., Savio, D., Farnleitner, A.H., Reischer, G.H., Ley, R.E., Mach, R.L., Farnleitner, A.H., Reischer, G.H., 2015. Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. Water Sci. Technol. 72, 1962–1972. https://doi.org/10.2166/wst.2015.407

Walters, W., Hyde, E.R., Berg-lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J. a, Jansson, J.K., 2015. Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. mSystems 1, e0009-15. https://doi.org/10.1128/mSystems.00009-15.Editor

Wang, Y., LêCao, K.-A., 2019. Managing batch effects in microbiome data. Brief. Bioinform. 00, 1–17. https://doi.org/10.1093/bib/bbz105

Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane, D.J., 1991. 16S Ribosomal DNA Amplification for Phylogenetic Study. J. Bacteriol. 173, 697–703.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R., 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5, 1–18. https://doi.org/10.1186/s40168-017-0237-y

Willis, A.D., 2019. Rarefaction, alpha diversity, and statistics. Front. Microbiol. 10. https://doi.org/10.3389/fmicb.2019.02407

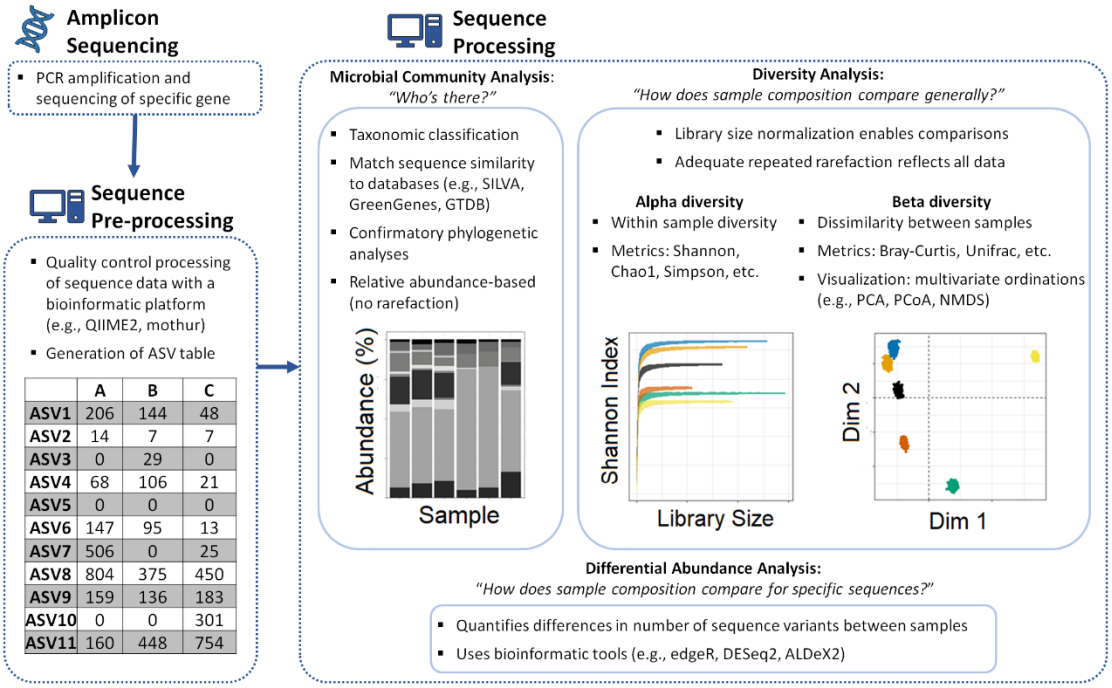
Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 87, 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

Yang, B., Wang, Y., Qian, P.Y., 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17, 1–8. https://doi.org/10.1186/s12859-016-0992-y

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O., 2014. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42, D643–D648. https://doi.org/10.1093/nar/gkt1209

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., Lu, Z., 2018. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Sci. Total Environ. 618, 1254–1267. https://doi.org/10.1016/j.scitotenv.2017.09.228

Zhang, L., Fang, W., Li, X., Lu, W., Li, J., 2020. Strong linkages between dissolved organic matter and the aquatic bacterial community in an urban river. Water Res. 184, 116089. https://doi.org/10.1016/j.watres.2020.116089



	A	B	C
ASV1	206	144	48
ASV2	14	7	7
ASV3	0	29	0
ASV4	68	106	21
ASV5	0	0	0
ASV6	147	95	13
ASV7	506	0	25
ASV8	804	375	450
ASV9	159	136	183
ASV10	0	0	301
ASV11	160	448	754

Figure 1: Schematic of general workflow in amplicon sequencing of water samples

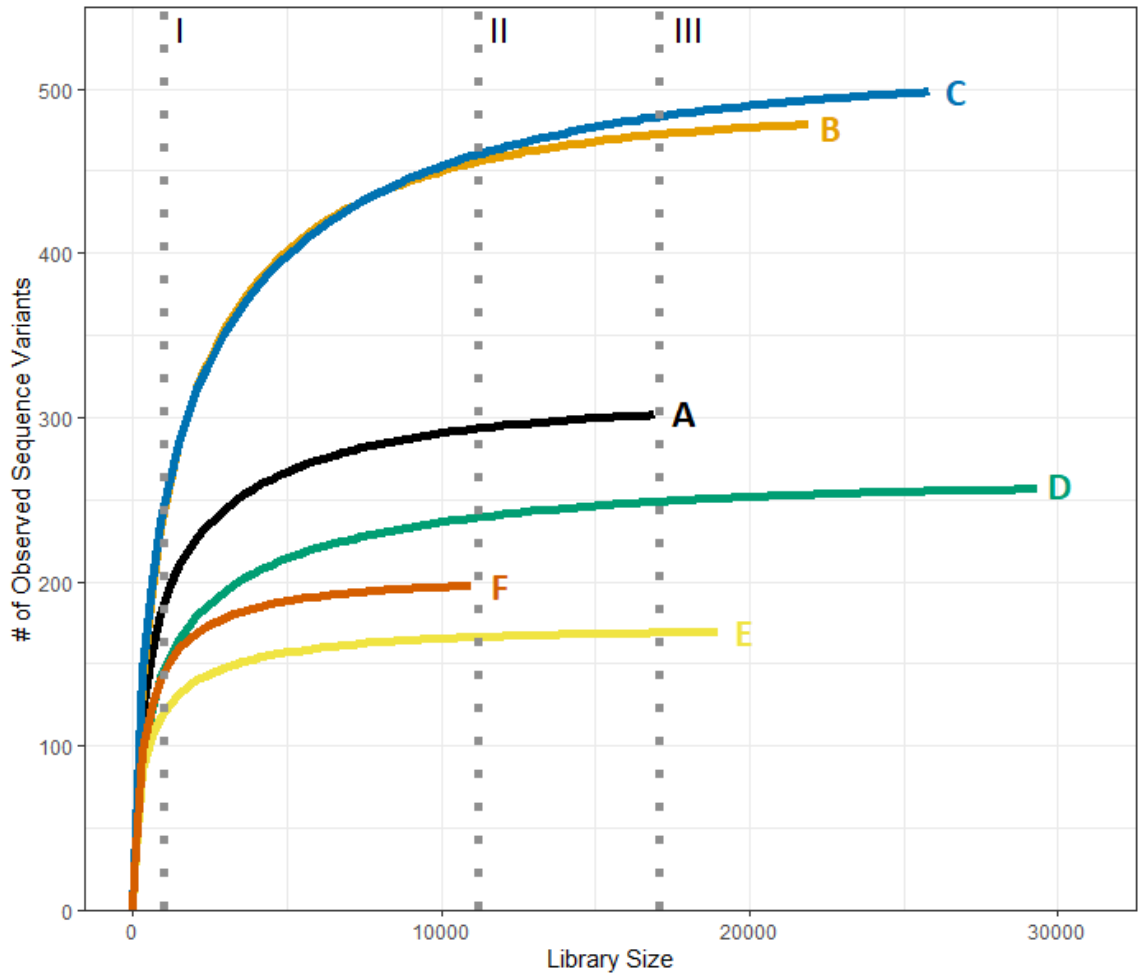


Figure 2: Rarefaction curves showing the number of unique sequence variants as a function of nomalized library size for six samples (labelled A – F) of varying diversity and intial library size. Selection of unnecessairily small library sizes (I) omits many sequences and potentially whole sequence variants. Rarefying to the smallest library size (II) omits fewer sequences and variants. While selection of a larger normalized library size (III) would omit even less sequences, it is necessary to omit entire samples (e.g., Sample F) that have too few sequences.

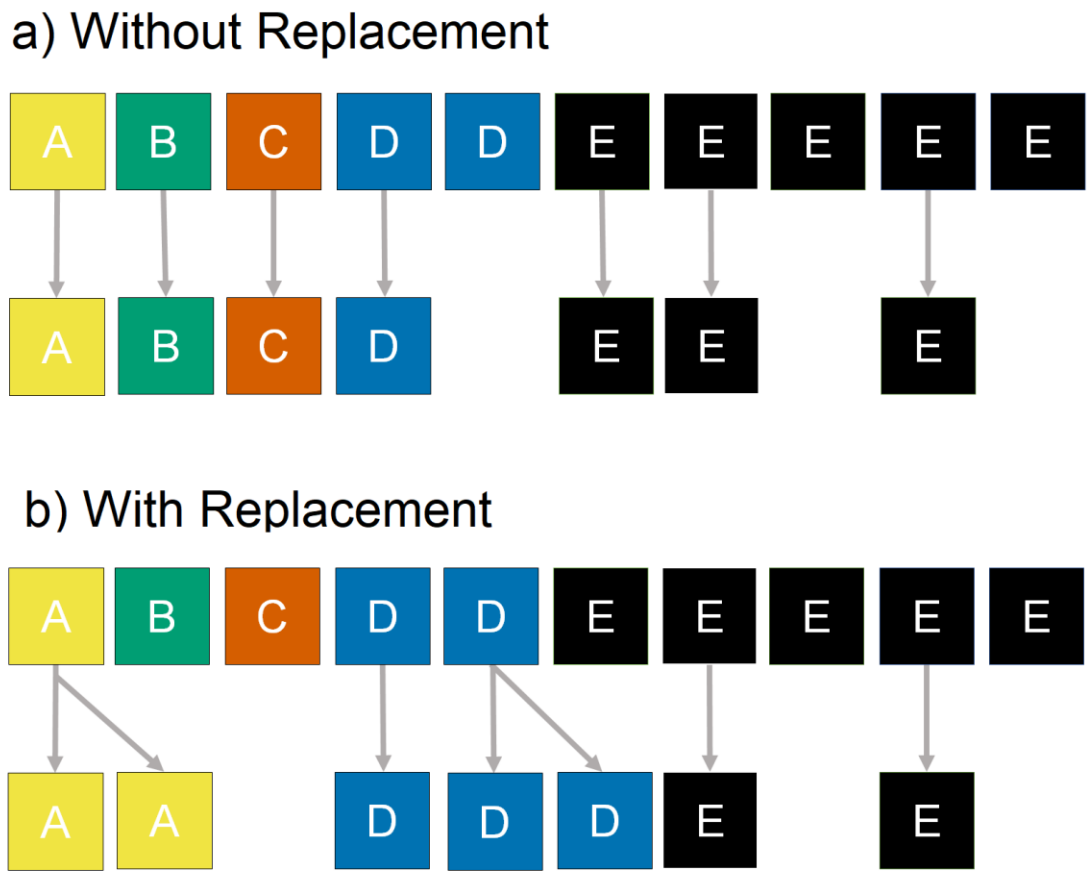


Figure 3: The mechanics of rarefying with or without replacement for a hypothetical sample with a library size of ten and five sequence variants (A – E). Rarefying without replacement (a) draws a subset from the observed library, while rarefying with replacement (b) has the potential to artificially inflate the numbers of some sequence variants beyond what was observed.

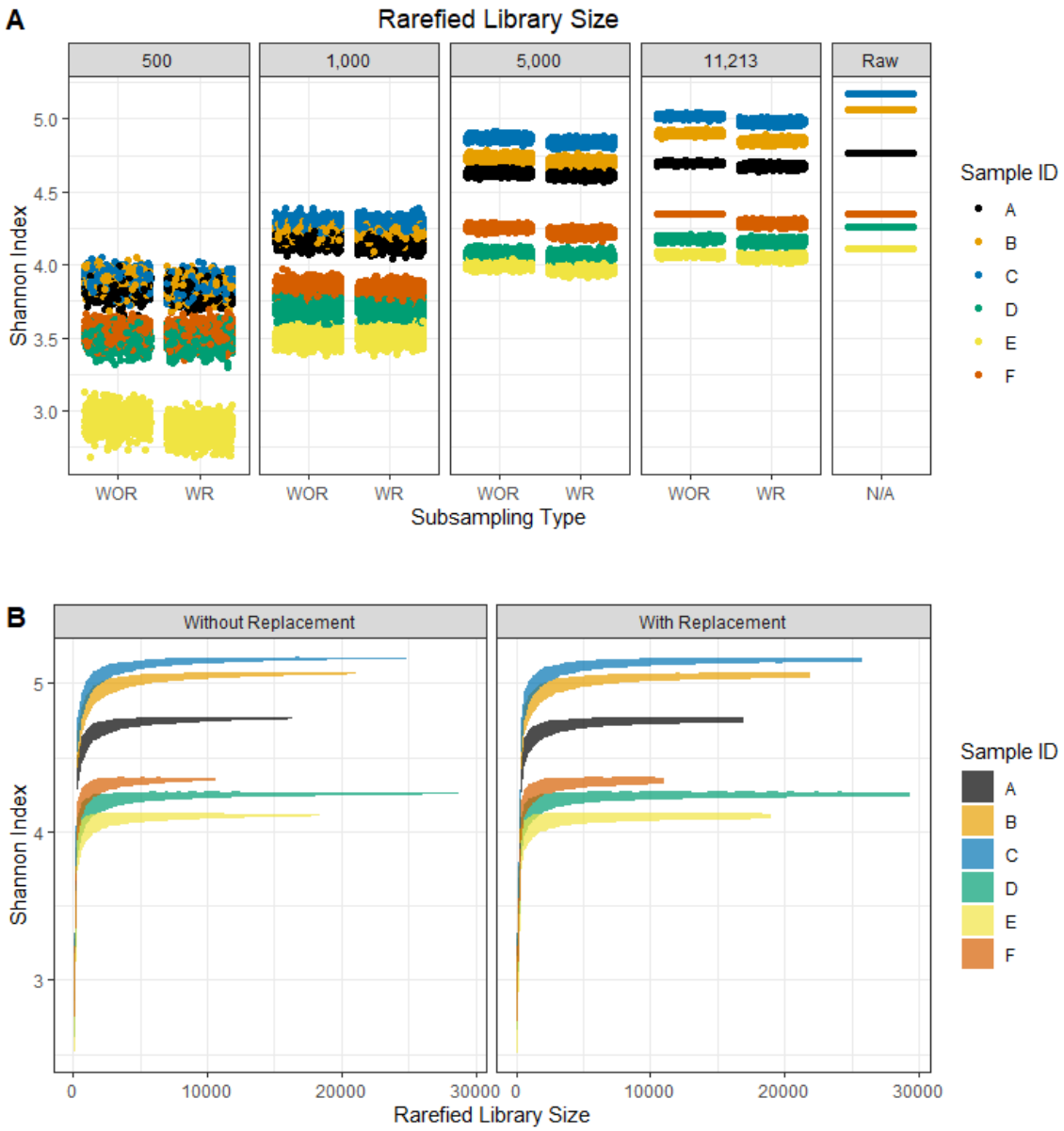


Figure 4: Effect of chosen rarefied library size and sampling with (WR) or without (WOR) replacement upon the Shannon Diversity Index. Six microbial communities were rarefied repeatedly (A) at specific rarefied library sizes of 11,213 sequences, 5,000 sequences, 1,000 sequences, and 500 sequences and (B) to evaluate the Shannon Index as a function of rarefied library size.

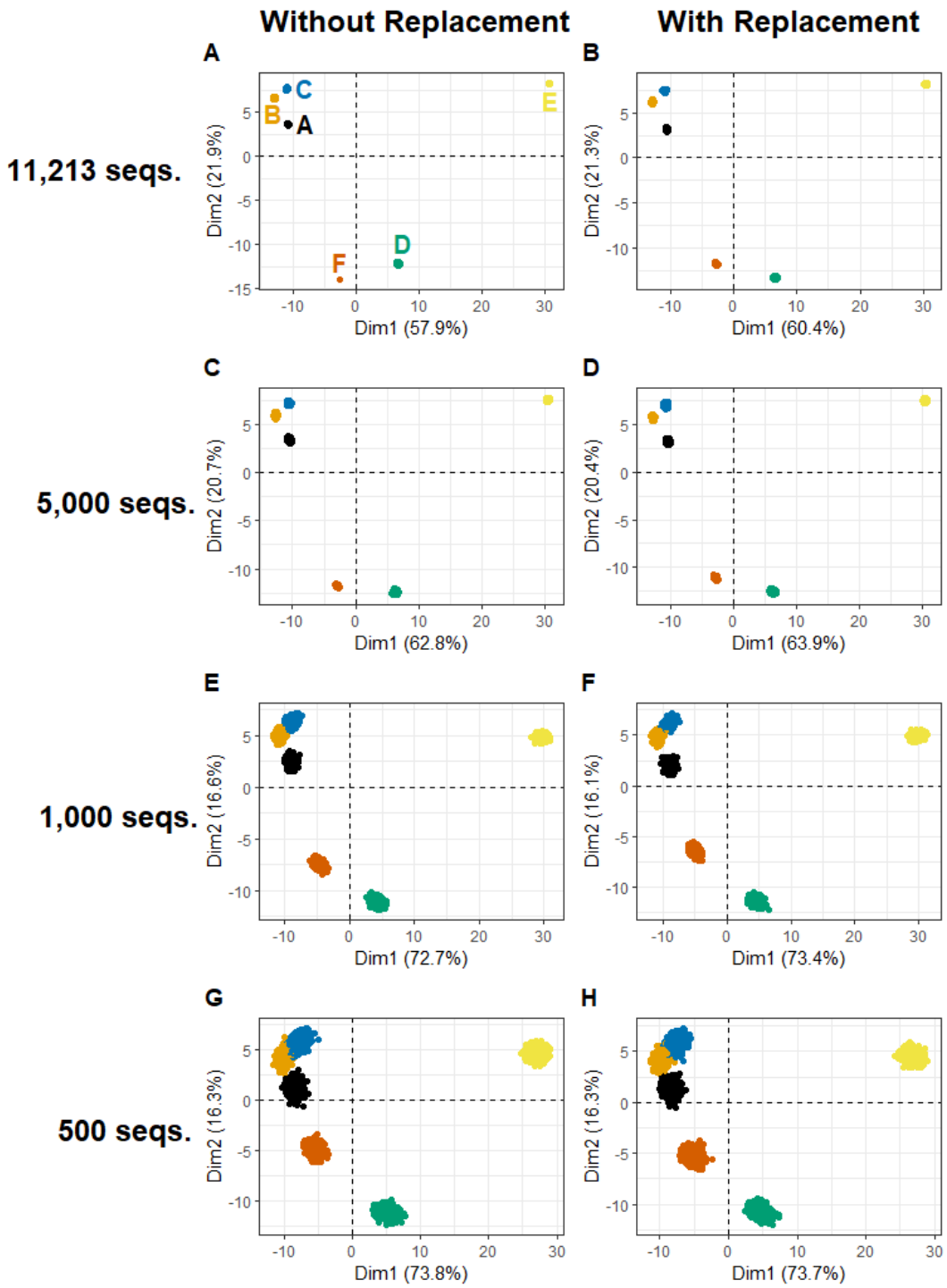


Figure 5: Variation in PCA ordinations (using the Bray-Curtis dissimilarity on Hellinger transformed rarefied libraries) of six microbial communities repeatedly rarefied with and without replacement to varying library sizes.

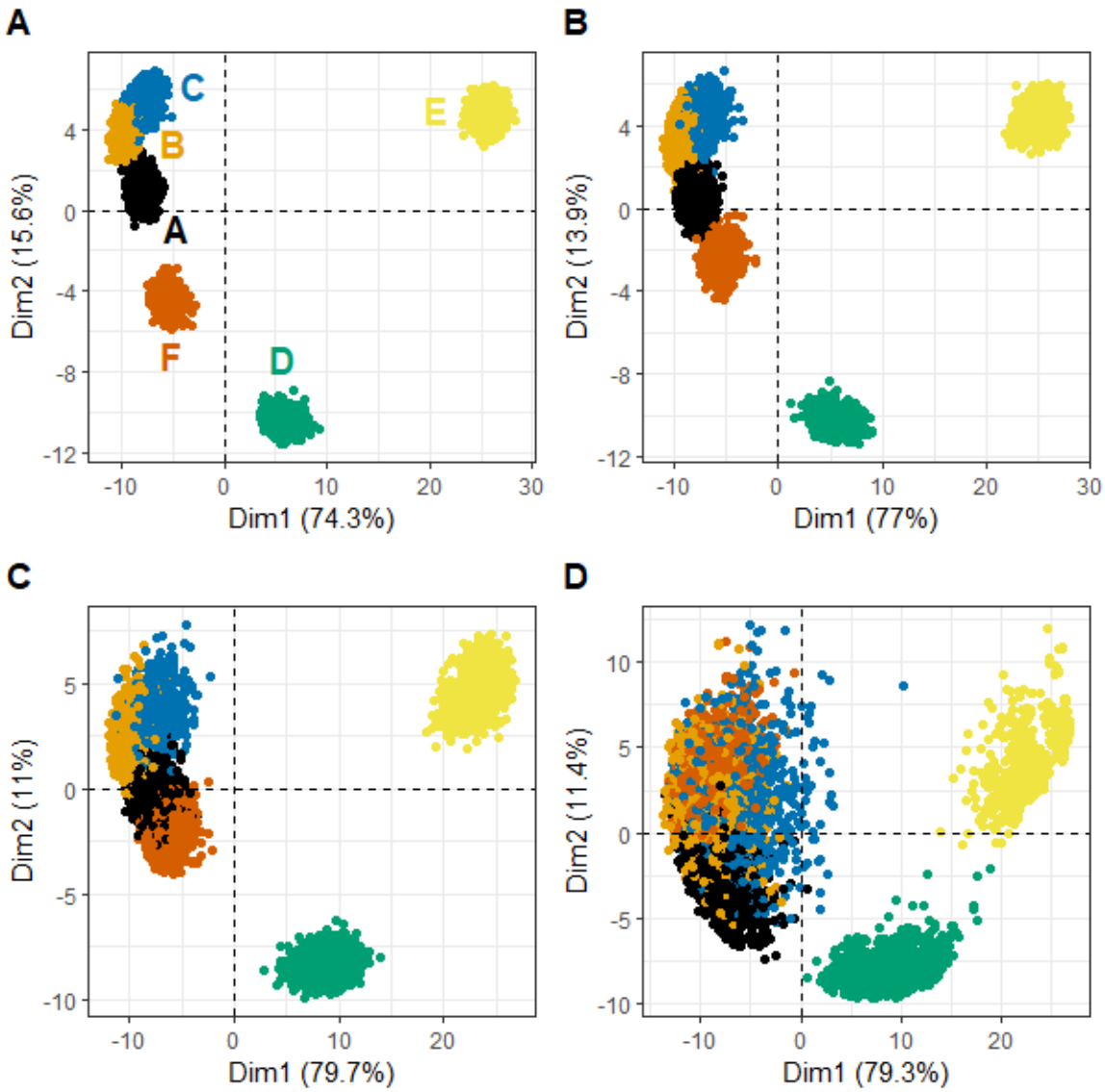


Figure 6: Variation in PCA ordinations (using the Bray-Curtis dissimilarity on Hellinger transformed rarefied microbial communities) of six microbial communities repeatedly rarefied to very small library sizes (A) 400, (B) 300, (C) 200 and (D) 100 sequences.

Table S1: Functions from other R packages used for *mirlyn* functionality

***mirlyn* function**	**Description**	**Functions used from other packages**	**Citations**
bartax()	Generate taxonomic composition barcharts from taxonomic abundance data.	microbiome::transform(): generate compositional data from abundance data.	Leo Lahti et al. microbiome R package. URL: http://microbiome.github.io
		phyloseq::tax_glom(): combine compositional data by desired taxonomic level. ggplot2::ggplot(): plotting engine for all visualization.	Paul J. McMurdie and Susan Holmes (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8(4):e61217.
alphawhichDF()	Calculate alpha diversity values from sequence count data.	vegan::diversity(): used for alpha diversity calculation.	Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan
alphacone()	Calculate alpha diversity values at different increments of library rarefaction for sequence data.	phyloseq::rarefy_even_depth(): rarefy taxonomic abundance data from multiple samples to an equal depth.	Paul J. McMurdie and Susan Holmes (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8(4):e61217.

		vegtan::diversity(): used for alpha diversity calculation.	Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan
betamatPCA()	Principle component analysis of beta diversity values calculated from sequence count data.	vegan::vegdist(): calculate dissimilarity indices from taxonomic abundance data.	Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan
		vegan::decostand(): apply desired standardization to taxonomic abundance data prior to calculation of dissimilarity indices.	
		stats::prcomp(): principle component analysis.	R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
mirl()	Repeated rarefaction of sequence count data.	phyloseq::rarefy_even_depth(): rarefy taxonomic abundance data from	Paul J. McMurdie and Susan Holmes (2013). phyloseq: An R package for reproducible

		multiple samples to an equal depth.	interactive analysis and graphics of microbiome census data. PLoS ONE 8(4):e61217.
rarefy_whole_rep()	Repeated rarefaction of taxonomic abundance data at incremental library sizes.	phyloseq::rarefy_even_depth(): rarefy taxonomic abundance data from multiple samples to an equal depth.	Paul J. McMurdie and Susan Holmes (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8(4):e61217.
rarecurve()	Visualize observed ASV count from rarefy_whole_rep() output.	ggplot2::ggplot(): plotting engine for all visualization.	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
alphawhichVis()	Visualize alpha diversity results from alphawhichDF() .	ggplot2::ggplot(): plotting engine for all visualization.	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
betamatPCAvis()	Plot principle component analysis from betamatPCA().	ggplot2::ggplot(): plotting engine for all visualization.	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
		factoextra::fviz_pca_ind(): wrapper for ggplot2 visualization of PCA data.	Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra