scientific reports

Check for updates

OPEN

Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities

Ellen S. Cameron[1], Philip J. Schmidt[2], Benjamin J.-M. Tremblay[1], Monica B. Emelko[2] & Kirsten M. Müller[1]✉

Amplicon sequencing has revolutionized our ability to study DNA collected from environmental samples by providing a rapid and sensitive technique for microbial community analysis that eliminates the challenges associated with lab cultivation and taxonomic identification through microscopy. In water resources management, it can be especially useful to evaluate ecosystem shifts in response to natural and anthropogenic landscape disturbances to signal potential water quality concerns, such as the detection of toxic cyanobacteria or pathogenic bacteria. Amplicon sequencing data consist of discrete counts of sequence reads, the sum of which is the library size. Groups of samples typically have different library sizes that are not representative of biological variation; library size normalization is required to meaningfully compare diversity between them. Rarefaction is a widely used normalization technique that involves the random subsampling of sequences from the initial sample library to a selected normalized library size. This process is often dismissed as statistically invalid because subsampling effectively discards a portion of the observed sequences, yet it remains prevalent in practice and the suitability of rarefying, relative to many other normalization approaches, for diversity analysis has been argued. Here, repeated rarefying is proposed as a tool to normalize library sizes for diversity analyses. This enables (i) proportionate representation of all observed sequences and (ii) characterization of the random variation introduced to diversity analyses by rarefying to a smaller library size shared by all samples. While many deterministic data transformations are not tailored to produce equal library sizes, repeatedly rarefying reflects the probabilistic process by which amplicon sequencing data are obtained as a representation of the amplified source microbial community. Specifically, it evaluates which data might have been obtained if a particular sample's library size had been smaller and allows graphical representation of the effects of this library size normalization process upon diversity analysis results.

Next-generation sequencing (NGS) has revolutionized the understanding of environmental systems by enabling characterization of microbial communities and their function through examination of DNA collected from samples that contain mixed assemblages of organisms[1–3]. It is well known that 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[4–6]. 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[2]. The use of NGS technology and techniques such as amplicon sequencing (sequencing of amplified genes of interest) have allowed for analysis of large quantities of diverse environmental DNA[7] and have largely eliminated challenges associated with culturing and microscopic identification[8] in this context. The relatively low cost of amplicon sequencing in comparison to other techniques (e.g., shotgun sequencing that sequences fragments of all present genetic material) has made it an increasingly popular technique[9,10].

The amplification and sequencing of specific genes (e.g., taxonomic marker genes) enables characterization of microbial community composition[11] ; as a result, it has been successfully applied in many areas of environmental and water research. This technique has been used to characterize and predict cyanobacteria blooms[12], describe

[1]Department of Biology, University of Waterloo, 200 University Ave. W, Waterloo, ON N2L 3G1, Canada. [2]Department of Civil and Environmental Engineering, University of Waterloo, 200 University Ave. W, Waterloo, ON N2L 3G1, Canada. ✉email: kirsten.muller@uwaterloo.ca

microbial communities found in aquatic ecosystems[13], and evaluate groundwater vulnerability to pathogen intrusion[14]. It has also been applied to water quality and treatment performance monitoring in diverse settings[15], including drinking water distribution systems[16,17], drinking water biofilters[18], anaerobic digesters[19], and cooling towers[20].

Processing and analysis of amplicon sequencing data are statistically complicated for a number of reasons[21]. In particular, 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 does not represent actual differences in microbial communities[8]. Amplicon sequencing libraries cannot be compared directly for this reason. For example, two replicate samples with 5,000 and 20,000 sequence reads, respectively, are likely to have different read counts for specific sequence variants simply due to the difference in library size. These differences in library size may arise due to differences in the amount of DNA loaded for sequencing and be further impacted by extraction efficiencies and primer bias during amplification, for example. While parametric tools such as generalized linear modelling[8] can provide a statistically sound framework for differential abundance analysis, drawing biologically meaningful diversity analysis conclusions from amplicon sequencing data typically requires normalization of library sizes. Such normalization accounts for the additional variation in counts that is attributable to differences in library sizes between samples[22]. For example, larger samples may appear more diverse than smaller samples[23] simply due to the presence of more sequences. 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[24]. This is accomplished by reducing the number of observations to a size threshold shared among several samples through random subsampling of the observations. Although initially developed for use in ecological studies, rarefaction is a commonly used library size normalization technique for amplicon sequencing data. It is, however, the subject of considerable debate and statistical criticism[8,25]. Rarefying is typically conducted in a single iteration that only provides a snapshot of the community that might have been observed at the smaller normalized library size. This introduces artificial variation to the data by omitting a random subset of observed sequences and potentially also necessitates discarding samples with library sizes deemed to be too small[8]. Alternatively, repeatedly rarefying has the potential to address the concerns associated with omission of data and could provide a more statistically acceptable technique than performing a single iteration of rarefying for diversity analyses. Nonetheless, rarefying repeatedly has received only trivial consideration in the literature[8,26], as discussed in further detail in the Background section. In concept, diversity analysis approaches grounded in statistical inference about source microbial diversity (that address the random probabilistic processes through which NGS yields libraries of sequence reads) could be superior to rarefying[27] , but they are not yet fully developed or readily available for routine diversity analysis to support study of environmental microbial communities.

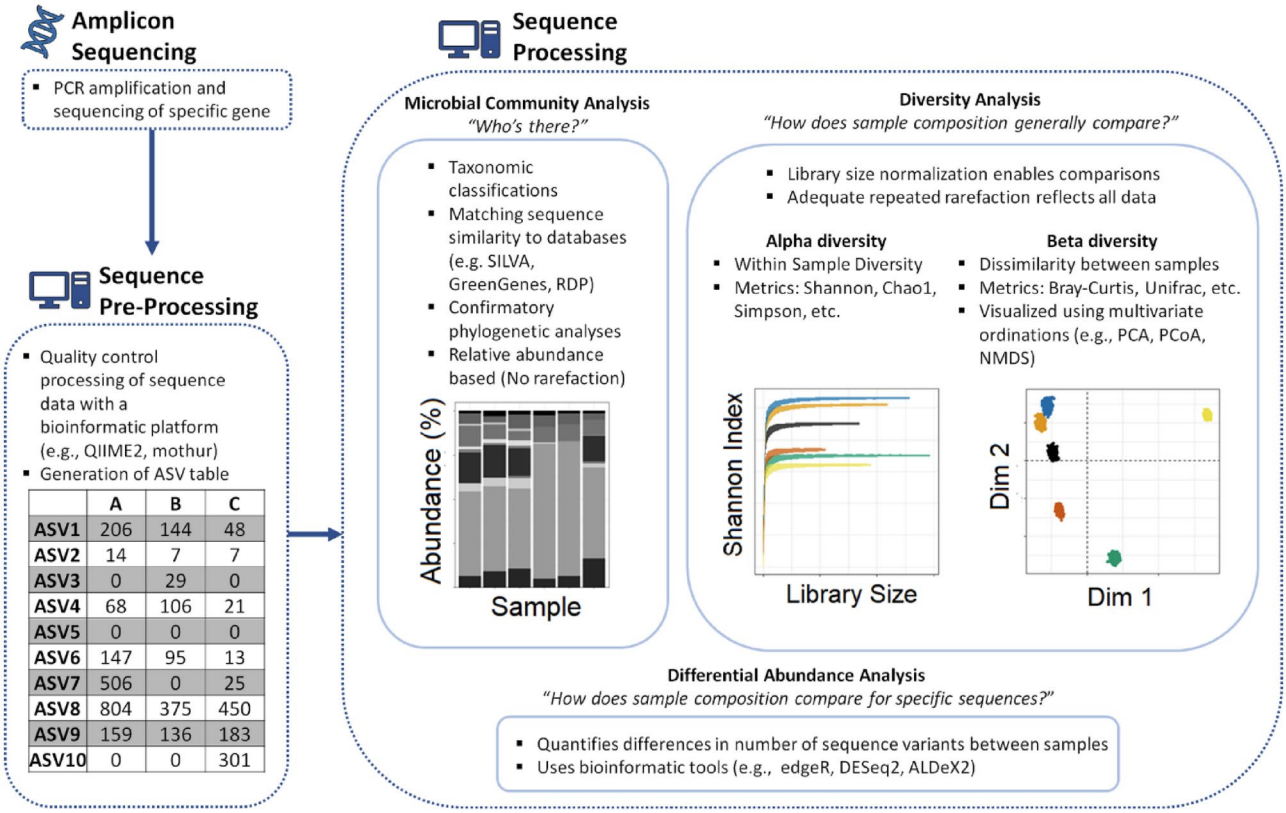
Here, we propose that rarefying repeatedly enhances assessment of the similarity or difference in diversity between samples by describing what data might have been obtained if a particular sample's library size had been smaller and characterizing the variability in diversity metrics introduced by rarefying samples to equal library sizes. Accordingly, application of repeatedly rarefying as a library size normalization technique is investigated in several illustrative diversity analyses. This paper graphically evaluates the impact of subsampling with or without replacement and normalized library size selection on diversity analyses such as the Shannon index and Bray–Curtis dissimilarity ordinations, specifically. Rather than representing diversity as a single numerical value or point in an ordination plot (often following transformation that may not be designed to compensate for differing library sizes), rarefying repeatedly yields bands of values or patches of points that characterize how diversity may vary among or between samples at a particular library size.

Background

Amplicon sequencing and diversity analysis for microbial communities in water: an overview. Due to the inevitable interdisciplinarity of environmental water quality research and the complexity and novelty of next generation sequencing relative to traditional microbiological methods used in water quality analyses, further background on amplicon sequencing, diversity analysis, and normalization is provided. Amplification and sequencing of taxonomic marker genes has been used extensively to examine phylogeny, evolution, and taxonomic classification of numerous groups across the three domains of life[28–30] . Taxonomic marker genes include the 16S rRNA gene in mitochondria, chloroplasts, bacteria and archaea[29,31–33], or the 18S rRNA gene within the nucleus of eukaryotes[34]. Widely used reference databases have been developed containing marker gene sequences across numerous phyla[2].

The 16S rRNA gene consists of nine highly conserved regions separated by nine hypervariable regions (V1–V9)[35] and is approximately 1540 base pairs in length[36,37] . While sequencing of the full 16S rRNA gene provides the highest taxonomic resolution[38] , many studies only utilize partial sequences due to limitations in read length of NGS platforms[36] which requires selection of an appropriate region of the 16S rRNA gene to amplify and sequence for optimal taxonomic resolution[36,39]. Sequencing the more conservative regions of the 16S rRNA gene may be limited to resolution of 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[33,36,39] . Different variable regions of the 16S rRNA gene may be biased towards different taxa[38] and be preferred for different ecosystems[40] making it critical to consider the suitability of the selected 16S rRNA region for the study area of interest.

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 in response to environmental conditions[11], and identification of differentially abundant taxa between samples[2]. Amplicon sequencing datasets



	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

Figure 1. Schematic of general workflow in amplicon sequencing of samples.

can be analyzed using a variety of bioinformatics pipelines for sequence analysis (e.g., sequence denoising, taxonomic classification, diversity analysis) including *mothur*[41] and *QIIME2*[42]. Such pipelines coordinate the outputs and inputs of various bioinformatic tools to ease the process of analyzing samples that require many different analysis steps. The implementation of denoising tools, such as *DADA2*[43] or *Deblur*[44], that are included in bioinformatic pipelines allows quality control of sequencing through the removal of sequencing errors and for the creation of amplicon sequence variants (ASVs), which retains the full observed biological variation by representing each unique sequence in a way that allows comparison between different studies[45]. Alternative methods for sequencing analysis may involve the creation of dataset-dependent operational taxonomic units (OTUs) through clustering sequences into groups that meet a certain similarity threshold[45].

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[21]. These counts, together with the total number of sequencing reads known as the library size[8], do not provide information on the absolute abundance of sequence variants[25,46]. This data can be used for studies on taxonomic composition, differential abundance analysis and diversity analyses (Fig. 1). Taxonomic classification of 16S rRNA sequences based on similarity to sequences in rRNA databases including SILVA[28] , the Ribosomal Database Project[47] and GreenGenes[48] allows for construction of taxonomic community profiles[1]. Taxonomic composition graphs frequently express community composition in proportions. Differential abundance analysis is utilized to explore whether specific sequence variants are found in significantly different proportions between samples[21] to identify potential biological drivers for these differences. This is frequently performed using programs initially designed for transcriptomics, such as *DESeq2*[49] and *edgeR*[50], or programs designed to account for the compositional structure of sequence data, such as *ALDeX2*[51]. The final potential application of this data is diversity analyses, which can be evaluated on varying scales from within sample (alpha) to between samples (beta)[52] but is associated with the challenge of the true diversity of environmental sources largely remaining unknown[53] .

Alpha diversity serves to identify richness (e.g., number of unique sequence variants observed) and evenness (e.g., allocation of read counts across observed sequence variants) within a sample[27]. Comparison of alpha diversity among samples of differing library sizes may result in inherent biases, with samples having larger library sizes appearing more diverse due to the potential presence of more sequence variants in samples with larger libraries[23,27]. This has commonly required samples to have equal library sizes before comparison to prevent bias fabricated only from differences in library size. Specific indices used to characterize the alpha diversity of samples include the Shannon index[54] , Simpson index[55], Chao1 index[56], and Hill number[57], but unique details of such indices should be understood for correct usage. For example, the Chao1 index relies on the observation of singletons in data to estimate diversity[56], but denoising processes for sequencing data may remove singleton reads, making the Chao1 estimator invalid for accurate analysis. The Shannon index used in this study is affected

by differing library sizes because the contribution of rare sequences to total diversity is progressively lost with smaller library sizes.

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 have more unique sequences simply due to the presence of more sequence variants[21]. A variety of beta diversity metrics can be used to compare sequence variant composition between samples including Bray–Curtis[58], Unifrac[59] or Jaccard[60] distances, which can then be visualized using ordination techniques (e.g., principal components analysis [PCA], principal coordinates analysis [PCoA], non-metric multidimensional scaling [NMDS]). Bray–Curtis dissimilarity, used in this study, includes pairwise comparison of the numbers for each ASV between two samples, and these are expected to be quite dissimilar (even if the communities they represent are not) if library sizes vary substantially.

Limitations of library size normalization techniques. Diversity analysis, as it is presently applied, usually requires library size normalization to account for bias introduced through varying read counts in samples. For example, samples with larger library sizes may appear more diverse simply due to the presence of more sequences. Normalization techniques that feature various statistical transformations have been proposed for use in place of rarefying or proportions[22], including upper-quartile log fold change[50], variance stabilizing transformations[49], centered log-ratio transformations[25], geometric mean pairwise ratios[61], or relative log expressions[62]. McKnight et al.[22] noted that the failure of most normalization techniques to transform data to equal library sizes for diversity analysis "is discouraging, as standardizing read depths are the initial impetus for normalizing the data (i.e., if all samples had equal read depths after sequencing, there would be no need to normalize".

These proposed alternatives to rarefying are also often compromised by the presence of large proportions of zero count data in tabulated amplicon sequencing read counts. Zero counts represent a lack of information[63] 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[64,65]. Nonetheless, many normalization procedures for amplicon sequencing datasets require zero counts to be omitted or modified, especially when applying transformations that utilize logarithms (e.g., upper-quartile log fold change, centered log-ratio, geometric mean pairwise ratios, relative log expressions). Methods that utilize logarithms involve fabricating count values (pseudocounts) for the many zeros of which amplicon sequencing datasets are often comprised and selecting an appropriate pseudocount value is an additional challenge[21] that may be accomplished using probabilistic arguments[25,46]. Zeros are a natural occurrence in discrete, count-based data such as the counting of microorganisms or amplicon sequences and adjusting or omitting them can introduce substantial bias into microbial analyses[66].

McMurdie and Holmes[8] noted that use of proportions is problematic due to heteroscedasticity: for example, one sequence read in a library size of 100 is a far less precise representation of source composition than 100 sequence reads in a library size of 10,000, even though both comprise 1% of the observed sequences. McKnight et al.[22] favour use of proportions in diversity analysis without noting how precision of proportions, and the degree to which alpha diversity in the source is reflected[27], varies with library size. Willis[27] points towards a conceptually better approach to diversity analysis that accounts for measurement error and the difference between the sample data and the population (environmental source) of which the sample data are only a partial representation. Diversity analysis in general does not do this, as it applies a set of calculations to sample data (or some transformation thereof) to obtain one value of alpha diversity or one point on an ordination plot. Pending further development of probabilistic approaches to diversity analysis[67], this study revisits rarefying because of the practical simplicity of comparing diversity among samples of equal library size.

McMurdie and Holmes [8] propose that rarefying is not a statistically valid normalization technique due to the omission of valid data. This may be resolved for the purposes of diversity analysis by rarefying repeatedly to represent all sequences in the proportions with which they were observed and compare sample-level microbial community diversity at a particular library size. McMurdie and Holmes[8] dismissed repeatedly rarefying as a normalization technique, in part because repeatedly rarefying an artificial library consisting of a 50:50 ratio of two sequence variants does not yield a 50:50 ratio at the rarefied library size and this added noise could affect downstream analyses. However, such error is inherent to subsampling, whether from a population or from a larger sequence library and has thus already affected samples with smaller library sizes; it is the reason why simple proportions are less precise in samples with smaller library sizes. Finally, McMurdie and Holmes[8], also cited the investigation of Navas-Molina et al.[26] as an example of repeatedly rarefying to normalize library sizes and used it to support their dismissal of this technique due to the omission of valid data and added variability. However, it is critical to note that the work in Navas-Molina et al.[26] reported using jackknife resampling of sequences, which cannot be equated to repeatedly rarefying (random resampling with or without replacement). Hence, it is necessary to build upon preliminary analysis of repeatedly rarefying as a normalization technique and to explore the impact of subsampling approach and normalized library size on diversity analysis results.

Methods

Example data—DNA extraction and amplicon sequencing. Samples used in this study are part of a larger study at Turkey Lakes Watershed (Ontario, Canada), but only an illustrative subset of samples is considered for the purpose of evaluating rarefaction rather than for ecological interpretation. This allows evaluation of repeated rarefying as a normalization technique without utilizing simulated data. DNA extracts isolated from environmental samples were submitted for amplicon sequencing using the Illumina MiSeq platform (Illumina Inc., San Diego, California) at the commercial laboratory Metagenom Bio Inc. (Waterloo, Ontario). Primers designed to target the 16S rRNA gene V4 region [515FB (GTGYCAGCMGCCGCGGTAA) and 806RB (GGA CTACNVGGGTWTCTAAT[68]) were used for PCR amplification.

Sequence processing and library size normalization. The program *QIIME2* (v. 2019.10)[42] was used for bioinformatic processing of sequence reads. Specifically, demultiplexed paired-end sequences were trimmed and denoised, including the removal of chimeric sequences and singleton sequence variants to avoid sequences that may not be representative of real organisms, using *DADA2*[43] to construct the ASV table. Zeroing all singleton sequences could erroneously remove legitimate sequences, particularly if the sequence in question is detected in large numbers in other similar samples; however, the potential effect of such error upon diversity analysis is beyond the scope of this work. Output files from *QIIME2* were imported into R (v. 4.0.1)[69] for community analyses using *qiime2R* (v. 0.99.23)[70]. Initial sequence libraries were further filtered using *phyloseq* (v. 1.32.0)[71] to exclude amplicon sequence variants that were taxonomically classified as mitochondria or chloroplast sequences.

We developed a package called *mirlyn* (Multiple Iterations of Rarefying for Library Normalization)[72] 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 (1) generate rarefaction curves, (2) repeatedly rarefy libraries to account for variation in library sizes among samples, and (3) plot diversity metrics given repeated rarefaction.

Community diversity analyses on normalized libraries. To demonstrate the impact of normalized library size on alpha diversity metrics, the Shannon index was evaluated. Further analyses using the Hill number demonstrate the conceptual application of this technique to other diversity metrics. The normalized libraries were also used for beta diversity analysis which was evaluated with the Bray–Curtis distance. 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[58]. Principal component analysis (PCA) was conducted on the Bray–Curtis distance matrices. Further analyses using Jaccard distances demonstrate the impact of rarefying repeatedly on metrics that reduce datasets to presence-absence composition.

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 explore the error associated with subsampling, samples were repeatedly rarefied 1000 times. This repetition provides a representative suite of rarefied samples capturing the randomness in sequence variant composition imposed by rarefying.

Rarefying library sizes may be performed with or without replacement. To evaluate the effects of subsampling replacement approaches, sequence libraries were repeatedly rarefied both 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.

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[8]. To evaluate the effects of different rarefied library sizes, sequence libraries were rarefied repeatedly to varying read depths. Results for various normalized library sizes were contrasted in diversity analyses to evaluate the impact of this determinant on interpretation of results.

Results and discussion

Use of rarefaction curves to explore suitable normalized library sizes. Suitable normalized library size for groups of samples can be determined through the examination of rarefaction curves (Fig. 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 reads 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., 1000 sequences) due to decreases in data quality[26]. 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 that is relevant to the study. 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 (Fig. 2I) or would require the omission of entire samples with smaller library sizes (Fig. 2III). While increasing sequencing depth would resolve the problem, this may not be an option for studies with limited resources or sample material. 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. Doing so with two or more normalized library sizes (e.g., an inclusive smaller value and a larger value that excludes some samples) may reveal differences in diversity in certain samples, particularly if effects of relatively rare variants are suppressed by normalizing to too small of a library size.

The effects of subsampling approach and normalized library size selection on alpha diversity analyses. 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 74. Sampling without replacement, however, is more statistically appropriate because it draws a subset from the observed set of sequences (as though the sample had yielded only the specified library size), whereas sampling with replacement fabricates a set of sequences in similar proportions to the observed set of sequences (Fig. 3). Sampling with

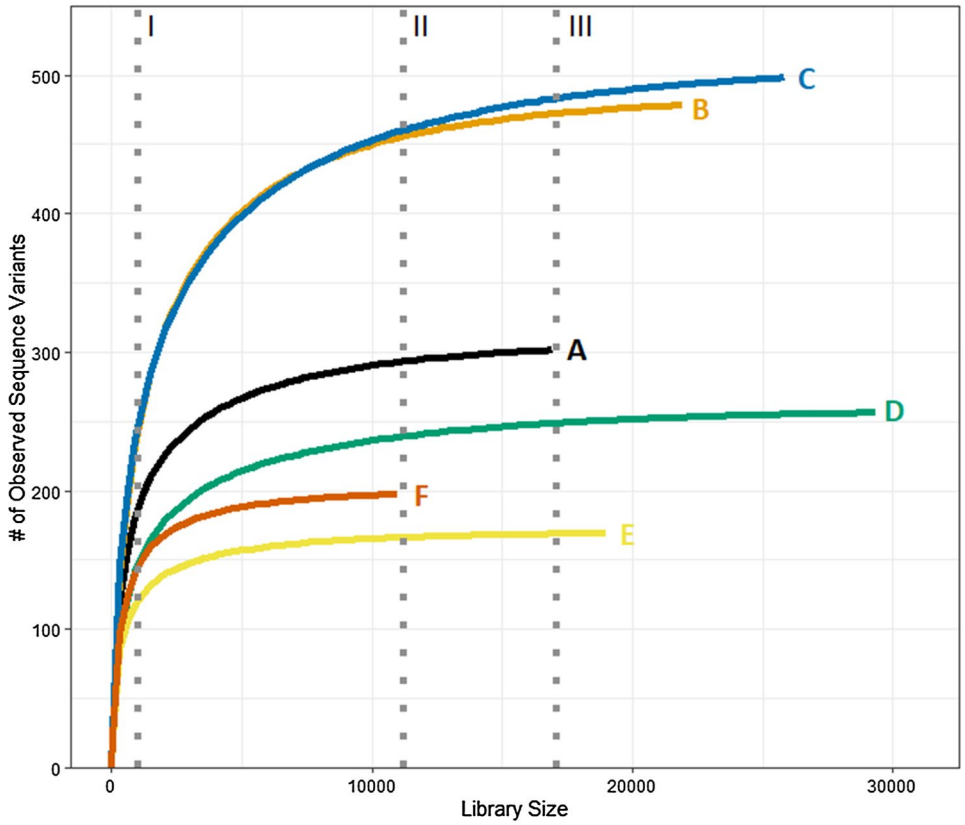


Figure 2. Rarefaction curves showing the number of unique sequence variants as a function of normalized library size for six samples (labelled A–F) of varying diversity and initial library size. Selection of unnecessarily small library sizes (I) omits many sequence variants. Rarefying to the smallest 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).

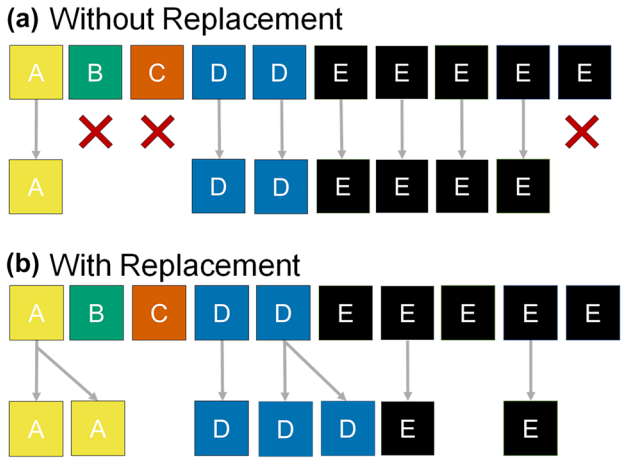


Figure 3. The mechanics of rarefying with or without replacement for a hypothetical sample with a library size of ten composed of five sequence variants (A–E). Rarefying without replacement (**a**) draws a subset from the observed library excluding the complementary subset, while rarefying with replacement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

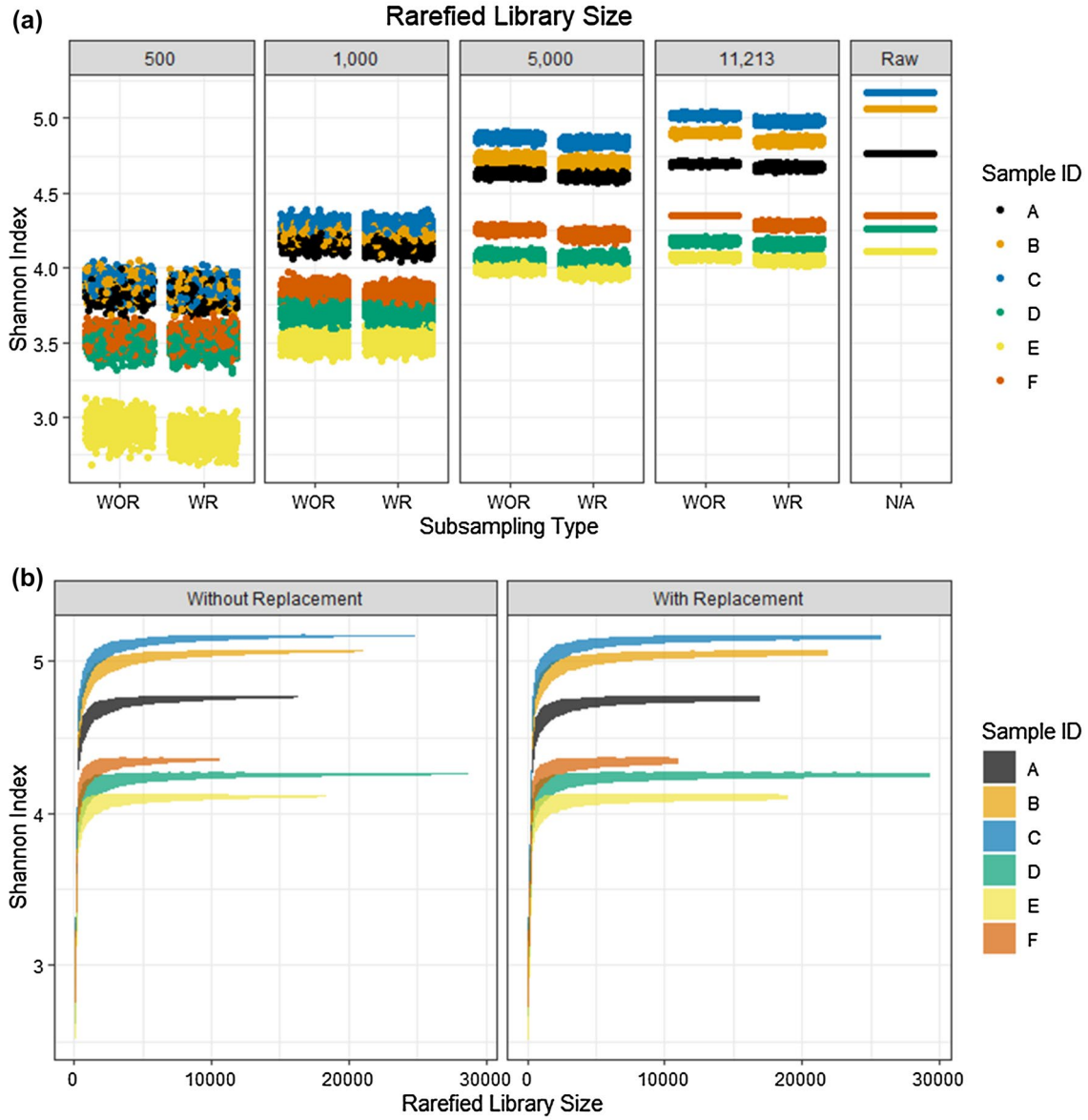


Figure 4. Effect of chosen rarefied library size and sampling with (WR) or without (WOR) replacement upon the Shannon Diversity Index. Six microbial communities were rarefied repeatedly (**A**) at specific rarefied library sizes of 11,213 sequences, 5000 sequences, 1000 sequences, and 500 sequences and (**B**) to evaluate the Shannon Index as a function of rarefied library size.

replacement can potentially cause a rare sequence variant to appear more frequently in the rarefied sample than it occurred 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 (Fig. 4A), 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 because rare sequences are excluded more often when sampling with replacement.

The conservation of larger normalized library sizes allows detection of more diversity with minimal variation observed between the iterations of rarefaction (Fig. 4A). The largest considered normalized library size (11,213, associated with the sample having the smallest library size) captured the highest Shannon index values, while the Shannon index diminishes for all samples at lower normalized library sizes. While there was only slight disparity in the Shannon index values between the largest rarefied 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 substantially reduced Shannon index value at a larger normalized library size. Alternatively, samples dominated by only a few highly abundant sequence variants will be comparatively robust to rarefying.

A plot of the Shannon index as a function of rarefied library size (Fig. 4B) 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 illustratively decreased to 5000, the Shannon index is still only slightly reduced by the rarefaction but there is greater variability introduced from rarefying.

The consistency of the diversity metric when rarefying repeatedly is extremely degraded when libraries were illustratively rarefied to the smallest considered rarefied library size of 500 sequences. This illustrates the potential to reach incorrect conclusions if rarefying is completed only once. When rarefying repeatedly to a small library size, however, diversity index values that are both highly inconsistent and suppressed relative to the diversity of the unrarefied data may lead to inappropriate claims of identical diversity values between samples (e.g., samples A, B, and C become indistinguishable). The extreme reduction and introduced variation of the Shannon index suggests that the selection of smaller rarefied library sizes (if necessary to include all samples) 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.

Similar trends were obtained when evaluating the Hill number instead of the Shannon Index (Figure S1-I), demonstrating the applicability of this concept to different diversity metrics. While similar trends were obtained when examining these data with a different diversity metric, in this case the Hill number, it is plausible that future analysis may reveal more distinctive differences with data featuring different sequence compositions.

Previous research evaluating normalization techniques has focused on beta diversity analysis and differential abundance analysis[8,21,25], but the appropriateness of library size normalization techniques for alpha diversity metrics should also be evaluated due to the prerequisite of having equal library sizes for calculation. Utilization of unnormalized library sizes with alpha diversity metrics may generate bias due to the potential for samples with larger library sizes to inherently reflect more of the diversity in the source than a sample with a small library size. The repeated iterations of rarefying library sizes allow characterization of the variability introduced to sample diversity by rarefying at any rarefied library size (Fig. 4) but (as is the case for all normalization-based approaches) does not allow evaluation of uncertainty about the diversity in the source from which the sample was taken.

The effects of subsampling approach and normalized library size selection on beta diversity analysis. When samples were repeatedly rarefied to a common normalized library size with and without replacement, similar amounts of variation in the Bray–Curtis PCA ordinations were observed between the sampling approaches (Fig. 5). 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 to represent data possible if only the smaller library size had been obtained, 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. For example, in the largest normalized library size possible for these data (Fig. 5A), a minimal amount of variation was 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 to retain as much information and introduce as little variation as possible. Although, a normalized library size of 5000 is lower than the flattening portion of the rarefaction curve for samples A, B, and C (Fig. 2), the selection of this potentially inappropriate normalized library size (Fig. 5C) can still accurately reflect the diversity between samples without excess artificial variation introduced through rarefaction. Due to the variation introduced to the Bray–Curtis dissimilarity ordinations in the smaller rarefied library sizes (Fig. 5E/G), it is critical to include computational replicates of rarefied libraries to fully characterize the introduced variation in communities (if such a small library size is needed to include all data). Notably, the dissimilarity between samples A and F diminishes with reduced normalized library sizes while the pattern of other samples persists (albeit with increasing variation introduced by rarefying). Similar trends were observed when using repeated rarefying with the Jaccard distance, showing the potential application of this technique with metrics focused on presence-absence data structures (Figure S1-II).

As discussed above, it has been suggested that repeatedly rarefying is inappropriate due to the introduction of "added noise". However, as demonstrated, repeatedly rarefying with larger rarefied library sizes is sometimes found to add only trivial variability to diversity analysis results, which is a useful outcome to defend the validity of library size normalization through rarefying. At smaller normalized library sizes, rarefaction without replication could result in artificial similarity or dissimilarity being identified between samples. Plotting the variability characterized by rarefying repeatedly aids the analyst in visually assessing similarity or dissimilarity of samples to avoid assertions that may be overly dependent on a single unusual rarefaction.

Beta diversity analysis of very small rarefied library sizes (Figure S2A, B, C) was performed to explore the robustness of these analyses and determine when the interpretation of the results would become severely impacted (Figure S2D). Repeatedly rarefying to extremely small library sizes can still reflect similar clustering patterns among samples observed in larger library sizes but with a much poorer resolution of clusters. Rarefying has previously been shown to be an appropriate normalization tool for samples with low sequence counts (e.g., < 1000 sequences per sample) by 21, which is promising for datasets containing samples with small initial library

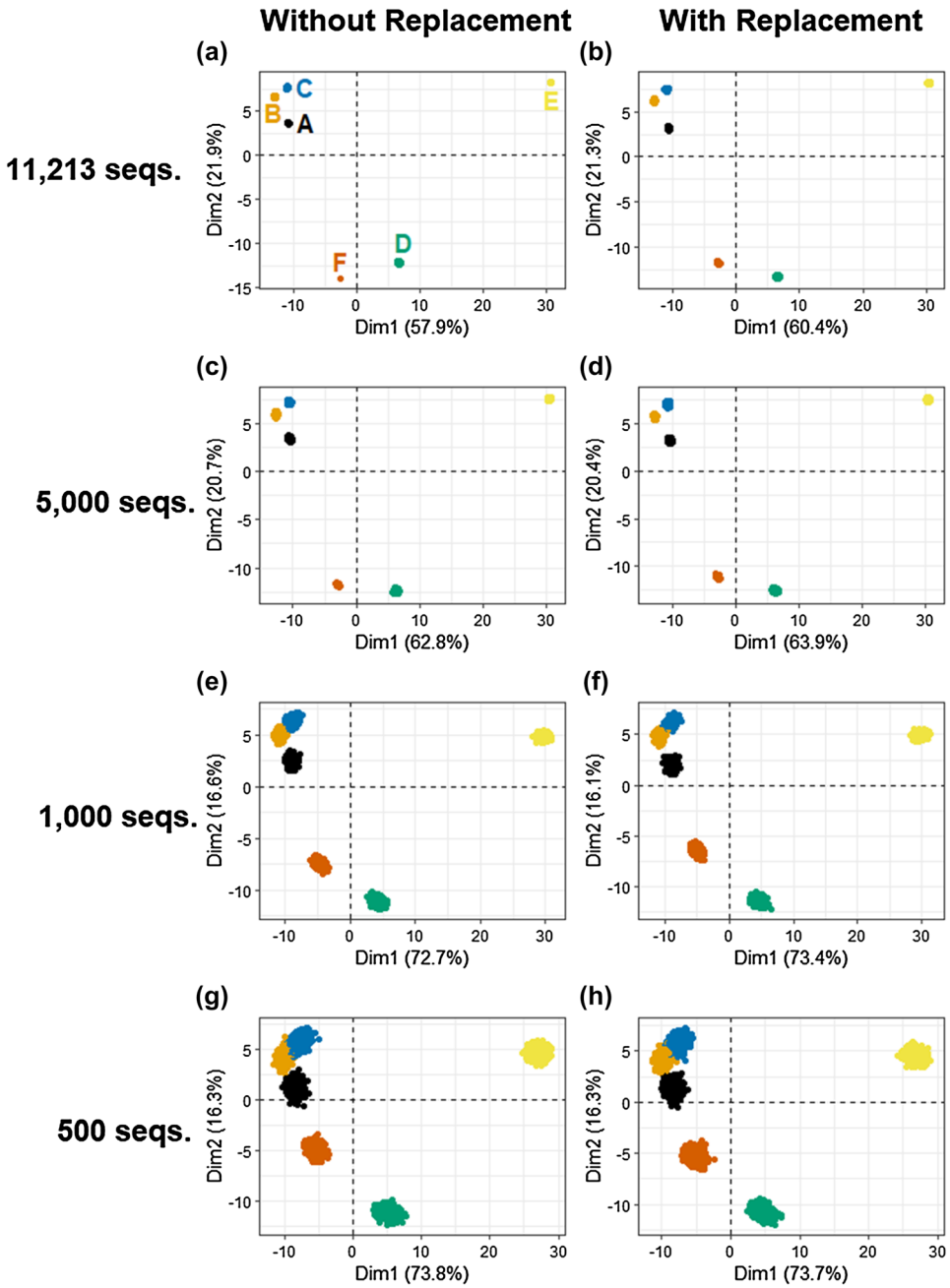


Figure 5. Variation in PCA ordinations (using the Bray–Curtis dissimilarity on Hellinger transformed rarefied libraries) of six microbial communities repeatedly rarefied with and without replacement to varying library sizes.

sizes or potentially analyzing subsets of data to explore diversity within specific phyla (e.g., Cyanobacteria). Caution must be taken to avoid selection of an excessively small normalized library size due to the introduction of an extreme level of artificial variation that compromises accurate depiction of diversity and suppresses the contribution of rare variants to overall diversity. The tradeoff between rarefying to a smaller than advisable library size or excluding entire samples with small library sizes 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 rarefying repeatedly does not become problematic until normalized library sizes are very small (e.g., 500 sequences or less) for the samples considered. While degradation of the consistency and value of the alpha diversity Shannon index at 500 sequences was observed, beta diversity analyses may be more robust to rarefaction and capable of reflecting qualitative clusters in ordination as previously discussed in Weiss et al.[21]. 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 to aid in determining if apparent similarity or dissimilarity is a chance result of rarefying.

The use of non-normalized 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[21]. However, the use of a single iteration of rarefying does result in the omission of valid data[8]. This study demonstrated that rarefying repeatedly and inclusion of these computational replicates in diversity analyses does not substantially impact the output and interpretation of beta diversity analyses unless rarefying to sizes that are inadvisably small to begin with. McMurdie and Holmes[8] were dismissive of rarefying repeatedly due to the variability it introduces, but such repetition was not evaluated in the context of beta diversity analysis. In the case of differential abundance analysis, the added variability of rarefying would be statistically inappropriate relative to generalized linear modelling that can account for varying library sizes.

McKnight et al.[22] preferred use of proportions in diversity analysis over rarefying (arguing that both were superior to other normalization approaches). While proportions normalize the sum of the ASV weights to one for each sample, the approach does not normalize the library size in terms of sequence counts. This is important because sample proportions will provide a more precise reflection of the true proportions of which the set of sequences is believed to be representative in samples with larger libraries than in samples with smaller libraries. In particular, using proportions of unnormalized sequence count libraries in beta diversity analysis overlooks the loss of alpha diversity associated with smaller library sizes when comparing samples with different library sizes.

Perspectives on library size normalization. The increasing popularity and accessibility of amplicon sequencing has enabled the scientific community to gain access to a wealth of microbial community data that would otherwise not have been accessible. However, despite amplicon sequencing of taxonomic marker genes being the gold standard approach for microbial community analysis, the data handling and statistical analysis is still in the early stages of development. The diversity analyses that the scientific community desires to perform on amplicon sequencing data require library sizes to be normalized across samples, which creates the challenge of determining appropriate normalization techniques. New normalization techniques and tools are constantly being developed and released to the community with claims that the newest technique is the best and only solution that should be utilized for analysis, but they may be associated with data handling limitations, be too specifically tailored to a particular type of analysis or desired property, or not normalize the library sizes that motivated the need for normalization [22] . For example, the centered-log ratio transformation [46] cannot be used with zero count data and amplicon sequencing datasets must be augmented with an artificial pseudocount to apply the normalization technique. The limitations of normalization techniques may affect downstream analyses, making it critical to understand the implications of the technique chosen.

Further discussion within the scientific community is needed to ensure rigorous interpretation of amplicon sequencing data without unwarranted bias introduced by the normalization technique. Approaches to microbiome data analysis that recognize data as samples from a source population and seek to draw inference about diversity in the source rather than just calculating diversity in the (transformed) sample are desirable. Random errors are inherent to sample collection, handling, processing, amplification, and sequencing and should ideally be reflected in how resulting data are analyzed[67]. Pending further research on such approaches, rarefying remains common in current research requiring library size normalization despite its potential limitations, especially for diversity analysis. The implementation of a single iteration of rarefying is problematic due to the omission of valid data and should not be used for library size normalization. Conducting repeated iterations of rarefying for the generation of a comprehensive collection of computational replicates for each sample, however, does not discard valid sequences and allows for the characterization of variation introduced through random subsampling in diversity analyses.

Library normalization remains a crucial step in diversity analyses, even with the increasing ability to generate samples with larger library sizes. If disparity exists between library sizes, a normalization technique is required to ensure that bias generated from data structure does not interfere with downstream analyses and subsequent interpretation. For example, if a threshold for library size is used in the generation of sequences (e.g., > 50,000 reads), disparity between samples may still exist when one sample may have 50,100 reads while a second has 75,000 reads. In these situations, normalization is still required, albeit for a relatively large normalized library size. Performing fewer repeated iterations may be suitable in some situations depending on the disparity between the initial library sizes and distribution of counts across different sequences. However, this can only be determined through initial data exploration and examination of the variation introduced through subsampling processes.

Rarefying and repeatedly rarefying have previously been generally discarded as library normalization techniques due to the omission of valid data, but this research proposes use of this technique for diversity analyses to encapsulate the variation introduced through rarefying. However, as is the case for initial criticisms of this technique, the results from this research may not be universally applicable and we caution that normalization techniques and data handling decisions should be driven by study design and the research question as there are certain cases where rarefying may not be an appropriate technique. For example, studies interested in the detection of very rare sequences may not benefit from this technique due to the increased probability of very rare sequences being regularly discarded. In these cases, researchers are encouraged to explore alternative data handling approaches to ensure rigorous analysis and prevent bias in the generation of results that are driven mainly by the initial data structure.

Conclusions

Repeated rarefying (e.g., 1000 times if computationally feasible) statistically describes possible realizations of the data if the number of sequences read had been limited to the normalized library size, thus allowing diversity analysis using samples of equal library size in a way that accounts for the data loss in rarefying.

Graphical depiction of the variability in diversity metrics introduced by rarefying allows analysts to make well-informed subject assessments and avert erroneous claims of similarity or dissimilarity that may arise from only rarefying once.

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 is theoretically more appropriate and will provide more accurate reflection of sample diversity.

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).

Bray–Curtis ordination patterns were 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 resulting in exclusion of the majority of the data (e.g., 100 sequences) is inappropriate because the substantial introduced variation leads to an inability to differentiate between sample clusters and suppresses contribution of rare variants to diversity.

Further development of strategies (e.g., data handling, library size normalization for diversity analyses) for ensuring rigorous interpretation of amplicon sequencing data is required.

Data availability

The datasets analyzed during this study are available for use as example data in the R package, *mirlyn* (https://github.com/escamero/mirlyn).

Received: 28 June 2021; Accepted: 27 October 2021
Published online: 16 November 2021

References

1. Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl. Environ. Microbiol.* **77**, 3846–3852 (2011).
2. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* **8**, 1561 (2017).
3. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21**, 1794–1805 (2012).
4. Bodor, A. *et al.* Challenges of unculturable bacteria: environmental perspectives. *Rev. Environ. Sci. Biotechnol.* **19**, 1–22 (2020).
5. Ferguson, R. L., Buckley, E. N. & Palumbo, A. V. Response of marine bacterioplankton to differential filtration and confinement. *Appl. Environ. Microbiol.* **47**, 49–55 (1984).
6. Cho, J. C. & Giovannoni, S. J. Cultivation and growth characteristics of a diverse group of oligotrophic marine gammaproteobacteria. *Appl. Environ. Microbiol.* **70**, 432–440 (2004).
7. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics: a guide from sampling to data analysis. *Microb. Inf. Exp.* **2**, 3 (2012).
8. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, (2014).
9. Clooney, A. G. *et al.* Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis. *PLoS ONE* **11**, 1–16 (2016).
10. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
11. Hodkinson, B. P. & Grice, E. A. Next-generation sequencing: a review of technologies and tools for wound microbiome research. *Adv. Wound Care* **4**, 50–58 (2015).
12. Tromas, N. *et al.* Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. *ISME J.* **11**, 1746–1763 (2017).
13. Zhang, L., Fang, W., Li, X., Lu, W. & Li, J. Strong linkages between dissolved organic matter and the aquatic bacterial community in an urban river. *Water Res.* **184**, 116089 (2020).
14. Chik, A. H. S. *et al.* Evaluation of groundwater bacterial community composition to inform waterborne pathogen vulnerability assessments. *Sci. Total Environ.* **743**, 140472 (2020).
15. Vierheilig, J. *et al.* Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. *Water Sci. Technol.* **72**, 1962–1972 (2015).
16. Perrin, Y., Bouchon, D., Delafont, V., Moulin, L. & Héchard, Y. Microbiome of drinking water: a full-scale spatio-temporal study to monitor water quality in the Paris distribution system. *Water Res.* **149**, 375–385 (2019).
17. Shaw, J. L. A. *et al.* Using amplicon sequencing to characterize and monitor bacterial diversity in drinking water distribution systems. *Appl. Environ. Microbiol.* **81**, 6463–6473 (2015).
18. Kirisits, M. J., Emelko, M. B. & Pinto, A. J. Applying biotechnology for drinking water biofiltration: advancing science and practice. *Curr. Opin. Biotechnol.* **57**, 197–204 (2019).
19. Lam, T. Y. C. *et al.* Superior resolution characterisation of microbial diversity in anaerobic digesters using full-length 16S rRNA gene amplicon sequencing. *Water Res.* **178**, 115815 (2020).
20. Paranjape, K. *et al.* Presence of *Legionella* spp. in cooling towers: the role of microbial diversity, Pseudomonas, and continuous chlorine application. *Water Res.* **169**, 115252 (2020).
21. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 1–18 (2017).
22. McKnight, D. T. *et al.* Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.* **10**, 389–400 (2019).
23. Hughes, J. B. & Hellmann, J. J. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods Enzymol.* **397**, 292–308 (2005).

24. Sanders, H. L. Marine benthic diversity : a comparative study. *Am. Nat.* **102**, 243–282 (1968).
25. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).
26. Navas-Molina, J. A. *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* **531**, 371–444 (2013).
27. Willis, A. D. Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* **10**, (2019).
28. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).
29. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S Ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173**, 697–703 (1991).
30. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579 (1990).
31. Case, R. J. *et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73**, 278–288 (2007).
32. Tsukuda, M., Kitahara, K. & Miyazaki, K. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs. *Sci. Rep.* **7**, 1–8 (2017).
33. Yang, B., Wang, Y. & Qian, P. Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinform.* **17**, 1–8 (2016).
34. Field, K. G. *et al.* Molecular phylogeny of the animal kingdom. *Science* **239**, 748–753 (1988).
35. Gray, M. W., Sankoff, D. & Cedergren, R. J. 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 (1984).
36. Kim, M., Morrison, M. & Yu, Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J. Microbiol. Methods* **84**, 81–87 (2011).
37. Schloss, P. D. & Handelsman, J. Status of the microbial census. *Microbiol. Mol. Biol. Rev.* **64**, 686–691 (2004).
38. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 1–11 (2019).
39. Bukin, Y. S. *et al.* The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci. Data* **6**, 1–14 (2019).
40. Escapa, I. F. *et al.* Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **8**, 65 (2020).
41. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
42. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
43. Callahan, B. J. *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
44. Amir, A. *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2:e00191**, e00191–16 (2017).
45. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
46. Gloor, G. B., Macklaim, J. M., Vu, M. & Fernandes, A. D. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian J. Stat.* **45**, 73–87 (2016).
47. Cole, J. R. *et al.* Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, 633–642 (2014).
48. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
51. Fernandes, A. D. *et al.* 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 (2014).
52. Sepkoski, J. J. Alpha, beta, or gamma: where does all the diversity go? *Paleobiology* (1988).
53. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**, 4399–4406 (2001).
54. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(369–423), 623–656 (1948).
55. Simpson, E. H. Measurment of diversity. *Nature* **163**, 688 (1949).
56. Chao, A. & Bunge, J. Estimating the number of species in a stochastic abundance model. *Biometrics* **58**, 531–539 (2002).
57. Jost, L. Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427–2439 (2007).
58. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).
59. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. USA.* **104**, 11436–11440 (2007).
60. Jaccard, P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. de la Soc. Vaudoise des Sci. Naturelles* **37**, 547–579 (1901).
61. Chen, L. *et al.* GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **2018**, 1–20 (2018).
62. Badri, M., Kurtz, Z., Muller, C. & Bonneau, R. Normalization methods for microbial abundance data strongly affect correlation estimates. *bioRxiv* 406264 (2018).
63. Silverman, J., Roche, K., Mukherjee, S. & David, L. Naught all zeros in sequence count data are the same. *bioRxiv* 477794 (2018).
64. Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* **26**, 330–335 (2016).
65. Wang, Y. & LêCao, K.-A. Managing batch effects in microbiome data. *Brief. Bioinform.* (2019).
66. Chik, A. H. S., Schmidt, P. J. & Emelko, M. B. Learning something from nothing: the critical importance of rethinking microbial non-detects. *Front. Microbiol.* **9**, 1–9 (2018).
67. Schmidt, P. J., Cameron, E. S., Müller, K. M. & Emelko, M. B. Ensuring that fundamentals of quantitative microbiology are reflected in microbial diversity analyses based on next-generation sequencing. *bioRxiv* (2021).
68. Walters, W. *et al.* Improved bacterial 16S rRNA Gene (V4 and V4–5) and Fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* **1**, e0009–15 (2015).
69. R Core Team. R: A language and environment for statistical computing. (2020).
70. Bisanz, J. E. qiime2R: Importing QIIME2 artifacts and associated data into R sessions. (2018).
71. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).
72. Cameron, E. S. & Tremblay, B. J.-M. mirlyn: Multiple iterations of rarefying for library normalization. (2020).

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.

Author contributions
E.S.C. wrote the draft of the manuscript, developed methodology and analysis. P.J.S. aided with conceptualization, writing (reviewing and editing), B.J.-M.T aided with methodology (development of R package). M.B.E. and K.M.M. supervised and aided with writing (reviewing and editing).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-01636-1.

Correspondence and requests for materials should be addressed to K.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.