frontiers in Microbiology

HYPOTHESIS AND THEORY
published: 01 March 2022
doi: 10.3389/fmicb.2022.728146

Ensuring That Fundamentals of Quantitative Microbiology Are Reflected in Microbial Diversity Analyses Based on Next-Generation Sequencing

*Philip J. Schmidt[1], Ellen S. Cameron[2], Kirsten M. Müller[2] and Monica B. Emelko[1]**

[1]Canada Research Chair in Water Science, Technology & Policy Group, Department of Civil and Environmental Engineering, Faculty of Engineering, University of Waterloo, Waterloo, ON, Canada, [2]Department of Biology, Faculty of Science, University of Waterloo, Waterloo, ON, Canada

OPEN ACCESS

Edited by:
Télesphore Sime-Ngando, Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Fernando Perez Rodriguez, University of Cordoba, Spain
Ammar Husami, Cincinnati Children's Hospital Medical Center, United States

***Correspondence:**
Monica B. Emelko
mbemelko@uwaterloo.ca

Specialty section:
This article was submitted to Aquatic Microbiology, a section of the journal Frontiers in Microbiology

Received: *21 June 2021*
Accepted: *20 January 2022*
Published: *01 March 2022*

Citation:
Schmidt PJ, Cameron ES, Müller KM and Emelko MB (2022) Ensuring That Fundamentals of Quantitative Microbiology Are Reflected in Microbial Diversity Analyses Based on Next-Generation Sequencing. Front. Microbiol. 13:728146. doi: 10.3389/fmicb.2022.728146

Diversity analysis of amplicon sequencing data has mainly been limited to plug-in estimates calculated using normalized data to obtain a single value of an alpha diversity metric or a single point on a beta diversity ordination plot for each sample. As recognized for count data generated using classical microbiological methods, amplicon sequence read counts obtained from a sample are random data linked to source properties (e.g., proportional composition) by a probabilistic process. Thus, diversity analysis has focused on diversity exhibited in (normalized) samples rather than probabilistic inference about source diversity. This study applies fundamentals of statistical analysis for quantitative microbiology (e.g., microscopy, plating, and most probable number methods) to sample collection and processing procedures of amplicon sequencing methods to facilitate inference reflecting the probabilistic nature of such data and evaluation of uncertainty in diversity metrics. Following description of types of random error, mechanisms such as clustering of microorganisms in the source, differential analytical recovery during sample processing, and amplification are found to invalidate a multinomial relative abundance model. The zeros often abounding in amplicon sequencing data and their implications are addressed, and Bayesian analysis is applied to estimate the source Shannon index given unnormalized data (both simulated and experimental). Inference about source diversity is found to require knowledge of the exact number of unique variants in the source, which is practically unknowable due to library size limitations and the inability to differentiate zeros corresponding to variants that are actually absent in the source from zeros corresponding to variants that were merely not detected. Given these problems with estimation of diversity in the source even when the basic multinomial model is valid, diversity analysis at the level of samples with normalized library sizes is discussed.

Keywords: amplicon sequencing, Shannon index, Markov chain Monte Carlo, normalization, rarefying

INTRODUCTION

Analysis of microbiological data using probabilistic methods has a rich history, with examination of both microscopic and culture-based data considered by prominent statisticians a century ago (e.g., Student, 1907; Fisher et al., 1922). The most probable number method for estimating concentrations from suites of presence–absence data is inherently probabilistic (e.g., McCrady, 1915), though routine use of tables (or more recently software) obviates consideration of the probabilistic link between raw data and the estimated values of practical interest for most users. Both the analysis of microbiological data and the control of the methods through which such data are obtained are grounded in statistical theory (e.g., Eisenhart and Wilson, 1943). More recently, the issue of estimating microbial concentrations and quantifying the uncertainty therein when some portion of microorganisms gathered in an environmental sample are not observed by the analyst has added to the complexity of analyzing microscopic enumeration data (e.g., Emelko et al., 2010). These examples share the common theme that the concentration of microorganisms in some source of interest is indirectly and imprecisely estimated from the discrete data produced by microbiological examination of samples (e.g., counts of cells/colonies or the number of aliquots exhibiting bacterial growth). The burgeoning microbiological analyses grounded in polymerase chain reactions (Huggett et al., 2015) likewise feature discrete objects (specific sequences of genetic material) that are prone to losses in sample processing, but these methods are further complicated by the variability introduced through amplification and subsequent reading (e.g., fluorescence signals or sequencing).

In next-generation amplicon sequencing, obtained data consist of a large library of nucleic acid sequences extracted and amplified from environmental samples, which are then tabulated into a set of counts associated with amplicon sequence variants (ASVs) or some grouping thereof (Callahan et al., 2017). The resulting data are regarded as a quantitative representation of the relative abundance (i.e., proportions) of various organisms in the source rather than absolute abundance (i.e., concentrations), thus leading to compositional data (Gloor et al., 2017). Among the many categories of analyses performed on such data are (1) differential abundance analysis to compare proportions of particular variants among samples and their relation to possible covariates and (2) diversity analysis that concerns the number of unique variants detected, how the numbers of reads vary among them, and how these characteristics vary among samples (Calle, 2019). Conventional analysis of these data is confronted with several problems (McMurdie and Holmes, 2014; Kaul et al., 2017; McKnight et al., 2019): (1) a series of samples can have diverse library sizes (i.e., numbers of sequence reads), motivating "normalization," (2) there are many normalization approaches from which to choose, and (3) many normalization and data analysis approaches are complicated by large numbers of zeros in ASV tables. These issues can be overcome in differential abundance analysis through use of probabilistic approaches such as generalized linear models (e.g., McMurdie and Holmes, 2014) that link raw ASV count data and corresponding library sizes to a linear model without the need for normalization or special treatment of zeros. Diversity analysis, however, is more complicated because the amount of diversity exhibited in a particular sample (alpha diversity) or apparent similarity or dissimilarity among samples (beta diversity) is a function of library size (Hughes and Hellmann, 2005), and methods to account for this are not standardized.

A variety of methods have been applied to prepare amplicon sequencing data for downstream diversity analyses, most of which involve some form of normalization. Normalization options include (1) rarefying that randomly subsamples from the observed sequences to reduce the library size of a sample to some normalized library size shared by all samples in the analysis (Sanders, 1968), (2) simple proportions (McKnight et al., 2019), and (3) a continually expanding set of data transformations, such as centered log-ratios (e.g., Gloor et al., 2017), geometric mean of pairwise ratios (e.g., Chen et al., 2018), or variance-stabilizing transformations (e.g., Love et al., 2014). Rarefying predates high throughput sequencing methods (including applications beyond sequencing of the 16S rRNA gene, such as RNA sequencing) and originated in traditional ecology. Statistically, these normalization approaches to estimation of sample diversity in the source treat manipulated sample data as a population because the non-probabilistic analysis of a sample (called a plug-in estimate) leads to a single diversity value or a single point on an ordination plot.

While it would increase computational complexity to do so, it is more theoretically sound to acknowledge that the observed library of sequence reads in a sample is an imperfect representation of the diversity of the source from which the sample was collected and that no one-size-fits-all normalization of the data can remedy this. ASV counts would then be regarded as a suite of random variables that are collectively dependent on the sampling depth (library size) and underlying simplex of proportions that can only be imperfectly estimated from the available data. Analysis of election polls is somewhat analogous in that it concerns inference about the relative composition (rather than absolute abundance) of eligible voters who prefer various candidates. A key distinction is that such analysis does not presume that the fraction of respondents favoring a particular candidate or party (or some numerical transformation thereof) is an exact measurement of the composition of the electorate. Habitual reporting of a margin of error with proportional results (Freedman et al., 1998) exemplifies that such polls are acknowledged to be samples from a population in which the small number of eligible voters surveyed is central to interpretation of the data. In amplicon sequencing diversity analysis, sampling precludes measuring source diversity exactly, but it is analysis of this source diversity that is of interest. Recognizing this, Willis (2019) encouraged approaches to estimation of alpha diversity from amplicon sequencing data that estimate diversity in a source and uncertainty therein from sample data using knowledge about random error. To do this, it is critical to recognize that diversity analysis of observed amplicon sequencing counts must draw inferences about source diversity without bias and with adequate precision.

Ideally, quantitative information about precision (such as error bars for alpha diversity values) should be provided.

Here, (1) the random process yielding amplicon sequencing data believed to be representative of microbial community composition in the source and (2) how this theory contributes to estimating the Shannon index alpha diversity metric using such data, particularly when library sizes differ and zero counts abound, are examined in detail. Theory applied to estimate microbial concentrations in water from data obtained using classical microbiological methods is extended to this type of microbiological assay to describe both the types of error that must be considered and a series of mechanistic assumptions that lead to a simple statistical model. The mechanisms leading to zeros in amplicon sequencing data and common issues with how zeros are analyzed in all areas of microbiology are discussed. Bayesian analysis is evaluated as an approach to drawing inference from a sample library about alpha diversity in the source with particular attention to the meaning and handling of zeros. This work addresses a path to evaluating microbial community diversity given the inherent randomness of amplicon sequencing data. It is based on established fundamentals of quantitative microbiology and provides a starting point for further investigation and development.

DESCRIBING AND MODELING ERRORS IN AMPLICON SEQUENCING DATA

A theoretical model for the error structure in microbial data can be developed by contemplating the series of mechanisms introducing variability to the number of a particular type of microorganisms (or gene sequences) that are present in a sample and eventually observed. This prioritizes understanding how random data are generated from the population of interest (e.g., the source microbiota) from sample collection through multiple sample processing steps to tables of observed data over the often more immediate problem of how to analyze a particular set of available data. Probabilistic modeling is central to such approaches, not just a data analytics tool.

Rather than reviewing and attempting to synthesize the various probabilistic methods that have been applied to amplicon sequencing, the approach herein builds on a foundation of knowledge surrounding random errors in microscopic enumeration of waterborne pathogens (e.g., Nahrstedt and Gimbel, 1996; Emelko et al., 2010) to address the inherently more complicated errors in amplicon sequencing data. This study addresses the foundational matter of inferring a source microbiota alpha diversity metric from an individual sample because dealing with more complex situations inherent to microbiome analysis requires a firm grasp of such simple scenarios. Accordingly, hierarchical models for alpha diversity analyses that link samples to a hypothetical meta-community (e.g., McGregor et al., 2020) and approaches for differential abundance analysis in which the covariation of counts of several variants among multiple samples may be a concern (e.g., Mandal et al., 2015) are beyond the scope of this work.

The developed modeling framework reflects that microorganisms and their genetic material are discrete, both in the source and at any point in the multi-step process of obtaining amplicon sequencing data. It also reflects that each step is random, potentially decreasing or in some cases increasing the (unobserved) discrete number of copies of each sequence variant. This applies a systems approach to describing the mechanisms through which discrete sequences are lost or created. This differs from previous work (e.g., McLaren et al., 2019) that does not reflect the discrete nature of microorganisms and their genetic material, assumes deterministic (i.e., non-random) and multiplicative effects of each sample processing step, and assumes only a single source of multiplicative random error in observed proportions (when, in fact, these proportions are estimated from observed discrete counts and a finite library size).

When random errors in the process linking observed data to the population characteristics of interest are integrated into a probabilistic model, it is possible to apply the model in a forward direction to simulate data given known parameter values or in a reverse direction to estimate model parameters given observed data (Schmidt et al., 2020). Analysis of simulated data generated in accordance with a mechanistically plausible probabilistic model (to determine if the analysis generates suitable inferences) is an important step in validating data-driven analysis frameworks. This reversibility is harnessed later in this paper to simulate data from a hypothetical source and evaluate how well Bayesian analysis of those data estimates the actual Shannon index of the source. The developed method is subsequently applied to a sample of environmental amplicon sequencing data.

Describing Amplicon Sequencing Data as a Random Sample From an Environmental Source

Microbial community analysis involves the collection of samples from a source, such as environmental waters or the human gut (Shokralla et al., 2012). This study addresses the context of water samples because the plausibility that some sources could be homogeneous provides a comparatively simple and well-understood statistical starting point for modeling—many other sources of microbial communities are inherently not well mixed. When a sample is collected, it is presumed to be representative of some spatiotemporal portion of a water source, such as a particular geographic location and depth in a water body and time of sampling. A degree of local homogeneity surrounding the location and time of the collected sample is often presumed so that randomness in the number of a particular type of microorganisms contained in the sample (random sampling error) would be Poisson distributed with mean equal to the product of concentration and volume. There are many reasons for which a series of samples presumed to be replicates from a particular source may yield microorganism counts that are overdispersed relative to such a Poisson distribution (Schmidt et al., 2014), including (1) clustering of microorganisms to each other or on suspended particles, (2) spatiotemporally

variable concentration, (3) variable volume analyzed, and (4) errors in sample processing and counting of microorganisms. Variable concentration and inconsistent sample volumes are not considered herein because the focus is on relative abundance (i.e., not estimation of concentrations) and samples that are not presumed to be replicates (i.e., analysis focuses on individual samples). Non-random dispersion could be a concern affecting estimates of diversity and relative abundance because clustering may inflate variability in the counts of a particular type of microorganisms. For example, clustering could polarize results between unusually large numbers if a large cluster is captured and absence otherwise rather than yielding a number that varies minimally around the average.

The remainder of this analysis focuses on errors in sample handling and processing, nucleic acid amplification, and gene sequence counting. To be representative of relative abundance of microorganisms in the source, it is presumed that a sample is handled so that the community in the analyzed sample is compositionally equivalent to the community in the sample when it was collected (Fricker et al., 2019). Any differential growth or decay among types of microorganisms or sample contamination will bias diversity analysis. A series of sample processing steps is then needed to extract and purify the nucleic material so that the sample is reduced to a size and condition ready for PCR. Losses may occur throughout this process, such as adhesion to glassware, residuals not transferred, failure to extract nucleic material from cells (Fricker et al., 2019), and sample partitioning during concentration and/or purification steps. These introduce random analytical error (because a method with 50% analytical recovery cannot recover 50% of one discrete microorganism, for example), and likely also non-constant analytical recovery if the capacity of the method to recover a particular type of microorganisms varies randomly from sample to sample (e.g., 60% in one sample and 40% in the next). Even with strict control of water matrix and method, the analytical recovery of microbiological methods is known to be highly variable in some cases (e.g., United States Environmental Protection Agency, 2005). Any differential analytical recovery among types of microorganisms (e.g., if one type of microorganisms is more likely to be successfully observed than another) will bias diversity analysis of the source (McLaren et al., 2019). Varying copy numbers of genes among types of microorganisms as well as genes associated with non-viable organisms can also bias diversity analysis if the goal is to represent diversity of viable organisms in the source rather than diversity of gene copies present in the source.

PCR amplification is then performed with specific primers to amplify targeted genes, which may not perfectly double the number of gene copies in each cycle due to various factors including primer match. Any differential amplification efficiency among types of microorganisms will bias diversity analysis of the source (McLaren et al., 2019), as will amplification errors that produce and amplify variants that do not exist in the source (unless these are readily identified and removed from sequencing data). Finally, the generated library of sequence reads is only a subsample of the sequences present in the amplified sample. Production of sequences that are not present in the original sample (e.g., chimeric sequences and misreads) is a form of loss if they detract from sequences that ought to have been read instead, and the resulting sequences may not be perfectly removed from the data (either failing to remove invalid sequences or erroneously removing valid sequences). Erroneous base calling is one such mechanism of sequencing error (Schirmer et al., 2015). Any differential losses at this stage will once again bias diversity analysis of the source (McLaren et al., 2019), as will inadvertent inclusion of false sequences. Thus, the discrete number of microorganisms gathered in a sample, the discrete number of genes successfully reaching amplification, the discrete number of genes after amplification, and the discrete number of genes successfully sequenced are all random. Due to this collection of unavoidable but often describable random errors, the validity of diversity analysis approaches that regard samples (or normalized transformations of them) as exact compositional representations of the source requires further examination.

Modeling Random Error in Amplicon Sequencing Data

For all of the reasons described above, it is impractical to regard libraries of sequence reads as indicative of *absolute* abundance in the source. We suggest that it is also impossible to regard them as indicative of *relative* abundance in the source without acknowledging a suite of assumptions and carefully considering what effect departure from those assumptions might have. By presuming that sequence reads are generated independently based on proportions identical to the proportional occurrence of those sequences in the source from which the sample was collected, the randomness in the set of sequence reads will follow a multinomial distribution (for large random samples from small populations, however, a multivariate hypergeometric model may be more appropriate). This is analogous to election poll data (if the poll surveys a small random sample of voters from a large electorate), repeatedly rolling a die, or repeatedly drawing random lettered tiles from a bag with replacement. This model may form the basis of logistic regression to describe proportions of sequences of particular types as a function of possible covariates in differential abundance analysis, reflecting how count data are random variables depending on respective library sizes and underlying proportions of interest.

Multinomial models are foundational to probabilistic analysis of count-based compositional data (e.g., McGregor et al., 2020), but mechanisms through which natural variability arises in the source (such as microorganism dispersion) and the sample collection and processing methodology (such as losses, amplification, and subsampling) must be considered because they may invalidate such a model for amplicon sequencing data—these need to be considered. **Table 1** summarizes the random errors discussed above, contextualizes them in terms of compatibility with the multinomial relative abundance model, and summarizes the assumptions that must be made to use a multinomial model. Although this table addresses the context of amplicon sequencing data, it could apply to other applications

TABLE 1 | Summary of random errors in amplicon sequencing and associated assumptions in the multinomial relative abundance model.

Error source	Description of error and compatibility with multinomial model	Assumptions
Sample collection	The random sampling error describing variability in the number of discrete objects captured in a sample yields a Poisson distribution if microorganisms are randomly dispersed in a large source. This error is compatible with a multinomial model for proportional abundance of variants. Clustering, including multiple gene copies per organism, leads to excess variability that is incompatible with a multinomial model.	• All microorganisms are randomly dispersed (i.e., not clustered) with only one gene copy each*
Sample handling	The number of a particular type of microorganisms may increase or decrease between sample collection and sample processing. Growth inflates the number of microorganisms at the level of diversity represented before growth occurred and is incompatible with a multinomial model. Decay is a form of random analytical error that is compatible with a multinomial model if it is consistent among variants.	• No growth • No differential decay (analytical recovery) among variants
Sample processing	The number of gene sequences subjected to amplification may be lower than the number in the sample prior to processing due to losses (e.g., adherence to apparatus, not all genes extracted, sample partitioning). This is compatible with a multinomial model if analytical recovery is constant among variants.	• No differential losses (analytical recovery) among variants
Amplification	The number of gene sequences is purposefully increased using polymerase chain reactions, inflating the number of gene sequences at the level of diversity represented before amplification occurred, and is incompatible with a multinomial model. Copy errors are a form of loss for the original sequences that were incorrectly copied and produces erroneous sequences that may then be further amplified. Erroneous sequences are incompatible with a multinomial model unless all of them are removed from the data.	• Pre-amplification variant diversity is fully identical to source diversity and sequences are perfectly duplicated in each PCR cycle* • No differential amplification efficiency or potential for copy errors among variants
Amplicon sequencing	Only a subsample of sequences are read, and all variants must be equally likely to be read. Sequence reading errors are a form of loss for the original sequences that were incorrectly read and also produces erroneous sequence reads. Sequence reading errors are incompatible with a multinomial model unless all resultant erroneous sequences are removed from the data.	• No differential sequence reading errors among variants or differential losses • Data denoising must remove all erroneous sequence reads and no legitimate reads

**Without these difficult assumptions, the multinomial model describes post-amplification variant diversity rather than source microbial diversity.*

of diversity analysis with specific modules excluded or modified (e.g., it could apply to metagenomics with amplification excluded for shotgun sequencing, and it may apply more broadly to ecology with modification).

Based on some simulations (see R code in **Supplementary Material**), it was determined that random sampling error consistent with a Poisson model is compatible with the multinomial relative abundance model (using the binomial model as a two-variant special case). Specifically, this featured Poisson-distributed counts of two variants with means following a 2:1 ratio and graphical evidence that this process is consistent with a binomial model (also with a 2:1 ratio of the two variants) when the result was conditioned on a particular library size. It must be noted that this is not a formal proof, as "proof by example" is a logical fallacy (unlike "disproof by counter-example"). Critically, clustering of gene copies in the source causes the randomness in sequence counts to depart from a multinomial model, as proven by simulation in the **Supplementary Material** (following a disproof by counter-example approach). When the above process was repeated with counts following a negative binomial model that is overdispersed with respect to the Poisson model, the variation in counts conditional on a particular library size was no longer consistent with the binomial model. Microorganisms having multiple gene copies is a form of clustering that invalidates the model.

Any form of loss or subsampling is compatible with the multinomial model so long as it affects all sequence variants equally. If each of a set of original proportions is multiplied by the same weight (analytical recovery), then the set of proportions adjusted by this weighting is identical to the original proportions (e.g., a 2:1 ratio is equal to a 1:0.5 ratio if all variants have 50% analytical recovery).

Growth and amplification must also not involve differential error among variants, but even in absence of differential error they have an important effect on the data and evaluation of microbiota diversity. These processes inflate the number of sequences present, but only with the potentially reduced or atypical diversity represented in the sample before such inflation. For example, a hypothetical sample with 100 variants amplified to 1,000 will have the diversity of a 100-variant sample in 1,000 reads, which may inherently be less than the diversity of a 1,000-variant sample directly from the source. Amplification fabricates additional data in a process somewhat opposite to discarding sequences in rarefaction; it draws upon a small pool of genetic material to make more whereas rarefaction subsamples from a larger pool of gene sequences to yield less (i.e., a smaller library size). Based on some simulations (see R code in **Supplementary Material**), it was proven that amplification is incompatible with the multinomial relative abundance model (following a disproof by counter-example approach). Specifically, the distribution of counts when two variants with a 2:1 ratio are amplified from a library size of four to a library size of six differs from the distribution of counts obtained from a binomial model.

Representativeness of source diversity and compatibility with the multinomial relative abundance model can only be assured

if the post-amplification diversity happens to be fully identical to the pre-amplification diversity and the observed library is a simple random sample of the amplified genetic material. Such an assumption may presume random happenstance more so than a plausible probabilistic process, though it would be valid in the extreme special case where pre-amplification diversity is fully identical to source diversity and every sequence is perfectly duplicated in each cycle (with no erroneous sequences produced). Without making relatively implausible assumptions or having detailed understanding and modeling of the random error in amplification, observed libraries are only representative of post-amplification diversity and indirectly representative of source diversity. This calls into question the theoretical validity of multinomial models as a starting point for inference about the proportional composition of microbial communities using amplification-based data. Nonetheless, the multinomial model was used as part of this study in some illustrative simulation-based diversity analysis experiments.

THE MANY ZEROS OF AMPLICON SEQUENCING DATA

As in other fields (Helsel, 2010), zeros in microbiology have led to much ado about nothing (Chik et al., 2018). They are (1) commonly regarded with skepticism that is hypocritical of non-zero counts (e.g., assuming that counts of zero result from error while counts of two are infallible), (2) often substituted with non-zero values or omitted from analysis altogether, and (3) a continued subject of statistical debate and special attention (such as detection limits and allegedly censored microbial data). Careful consideration of zeros is particularly relevant to diversity analysis of amplicon sequencing data because they often constitute a large portion of ASV tables. They may or may not appear in sample-specific ASV data, but they often appear when the ASV table of several samples is filled out (e.g., when an ASV that appears in some samples does not appear in others, zeros are assigned to that ASV in all samples in which it was not observed). They may also be created by zeroing singleton reads (Callahan et al., 2016), but this issue (and the bias arising if some singletons are legitimate read counts) is not specifically addressed in this study. Zeros often receive special treatment during the normalization step of compositional microbiome analysis (Thorsen et al., 2016; Tsilimigras and Fodor, 2016; Kaul et al., 2017), including removal of rows of zeros and fabrication of pseudo-counts with which zeros are substituted (to enable logarithmic transformations or optimal beta diversity separation). However, it is fundamentally flawed to justify use of pseudo-counts to "correct for" zeros arguing that they are censored data below some detection limit (Chik et al., 2018; Cameron et al., 2021). In methods that count discrete objects (e.g., microorganisms or gene copies) in samples, counts of zero are no less legitimate than non-zero counts: all random count-based data provide imperfect estimation of source properties, such as microorganism concentrations or proportional composition.

We propose a classification of three types of zeros: (1) non-detected sequences (also called rounded or sampling zeros), (2) truly absent sequences (also called essential or structural zeros), and (3) missing zeros. This differs from the three types of zeros discussed by Kaul et al. (2017) because the issue of missing zeros (which is shown to be critically important in diversity analysis) was not noted in that study and zeros that appear to be outliers from empirical patterns are not considered in this study (because all random read counts are presumed to be correct). Likewise, Tsilimigras and Fodor (2016) differentiated between essential or structural zeros (truly absent sequences) and rounded zeros (resulting from undersampling), but did not consider missing zeros.

It is typically presumed that zeros correspond to non-detected sequences, meaning that the variant is present in some quantity in the source but happened to not be included in the library and is represented by a zero. A legitimate singleton that is replaced with a zero would be a special case of a non-detect zero. Bias would result if non-detect zeros were omitted or included in the diversity analysis inappropriately (e.g., substitution with pseudo-counts or treating them as definitively absent variants). It is conceptually possible that a particular type of microorganisms may be truly absent from certain sources so that the corresponding read count and proportion should definitively be zero. If false sequences due to errors in amplification and sequencing are filtered from the ASV table but left as zeros, then they are a special case of truly absent sequences. Bias would result if such zeros were included in diversity analysis in a way that manipulates them to non-zero values or allows the corresponding variant to have a plausibly non-zero proportion. Missing zeros are variants that are truly present in the source and not represented in the data—they are not acknowledged to be part of the community, even with a zero in the ASV table. Bias would result from excluding these zeros from diversity analysis rather than recognizing them as non-detected variants. Thus, there are three types of zeros, two of which appear indistinguishably in the data and must be handled differently and the third of which is important but does not even appear in the data. In this study, simulation-based experiments and environmental data are used to illustrate implications of the dilemma of not knowing how many zeros should appear in the data to be analyzed as non-detects.

PROBABILISTIC INFERENCE OF SOURCE SHANNON INDEX USING BAYESIAN METHODS

The Shannon index (Shannon, 1948; Washington, 1984) is used as a measure of alpha diversity that reflects both the richness and evenness of variants present (number of unique variants and similarity of their respective proportions). When calculated from a sample, the Shannon index (S) depends only on the proportions of the observed variants (p_i for the ith of n variants) and not on their read counts. Critically, the Shannon index of a sample is not an unbiased estimate of the Shannon index of the source (even in scenarios without amplification); it is

expected to increase with library size as more rare variants are observed until it converges asymptotically on the Shannon index of the source (Willis, 2019). Even if all variants in the source are reflected in the data, the precision of the estimated Shannon index will improve with increasing library size.

$$S = -\sum_{i=1}^{n} p_i \times \ln p_i$$

Building on existing work applying Bayesian methods to characterize the uncertainty in enumeration-based microbial concentration estimates (e.g., Emelko et al., 2010) and inspired by the need to consider random error in evaluation of alpha diversity that was noted by Willis (2019), a Bayesian approach is explored here for the simplified scenario of multinomially distributed data. It evaluates uncertainty in the source Shannon index given sample data, the multinomial model, and a relatively uninformative Dirichlet prior that gives equal prior weight to all variants (using a vector of ones). Hierarchical modeling that may describe how the proportional composition varies among samples is beyond the scope of this analysis. Such modeling can be beneficial when strong information in the lower tier of the hierarchy can be used to probe the fit of the upper tier; however, it can be biased if limited information in the lower tier is bolstered with flawed assumptions introduced *via* the upper tier.

Here, a simulation study is employed that is analogous to compositional microbiota data with small library sizes and small numbers of variants and that does follow a multinomial relative abundance model. The simulation uses specified proportions for a set of variants; for illustrative purposes, the simulation represents random draws with replacement from a bag of lettered tiles based on the game Scrabble™. Randomized multinomial data (**Supplementary Table S1**, **Supplementary Material**) were generated in R using varying library sizes and the proportions of the 100 tiles (including 26 letters and blanks), which correspond to a population-level Shannon index of 3.03. Markov chain Monte Carlo (MCMC) was carried out using OpenBUGS (version 3.2.3), with randomized initialization and 10,000 iterations following a 1,000-iteration burn-in. The model specification code and a small sample dataset are included in the **Supplementary Material**. Due to the mathematical simplicity of a multinomial model with a Dirichlet prior, this number of iterations can be completed in seconds with rapid convergence and good mixing of the Markov chain. Each iteration generates an estimate of each variant proportion, and the set of variant proportions is used to compute an estimate of the Shannon index for the source inferred from the sample data. The Markov chain of Shannon index values generated in this way is collectively representative of a sample from the posterior distribution that characterizes uncertainty in the source Shannon index given the sample data and prior. The simulated data were analyzed in several ways, as illustrated using box and whisker plots in **Figure 1**: (1) with all non-detected tile variants removed, (2) with zeros added as needed to reach the correct number of tile variants used to simulate the data (i.e., 27), and (3) with extraneous zeros (a total of 50 tile variants of which 23 do not actually exist in the source).

The disparity in results between the three ways in which the data were analyzed exemplifies the importance of zeros in estimating the Shannon index of the source from which a sample was gathered. Omitting non-detect zeros in this Bayesian analysis characteristically underestimates diversity, while including zeros for variants that do not exist in the source characteristically overestimates diversity. In each case, the effect diminishes as the library size is increased. Notably, the approach that included only zeros for variants present in the source that were not detected in the sample allowed accurate estimation of the source Shannon index, with improving precision as the library size increases (exemplifying statistical consistency of the estimation process). Given these results, the proposed Bayesian process appears to be theoretically valid to estimate the source Shannon index from samples (for which the multinomial relative abundance model applies), and it does so without the need to normalize data with differing library sizes. Practically, however, it is not possible to know how many zeros should be included in the analysis estimating the Shannon index because the number of unique variants actually present in the source is unknown. This is a peculiar scenario that must be emphasized here because accurate statistical inference about the source is impossible: although the model form (multinomial) is known, the number of unique variants that should be included in the model is practically unknowable. Model-based supposition is not applied in this study to introduce information that is lacking; this can be a biased approach to compensating for deficiencies in observed data or flawed experiments in which "control variables" are not controlled (e.g., it is not possible to estimate concentration from a count without a measured volume) unless the supposition happens to be correct (Schmidt et al., 2020).

Because the extent to which zeros compromised accurate estimation waned with increasing library size (**Figure 1**), a similar analysis was performed on amplicon sequencing data for six water samples from lakes. The samples (Cameron et al., 2021, **Supplementary Data Sheet 2**) featured library sizes between 10,000 and 30,000 and observation of 1,142 unique variants among the samples. All singleton counts had been zeroed and the completed ASV table had 3,342 rows (2,200 of which are all zeros associated with variants detected in other samples from the same study area). Each sample was analyzed three ways: (1) with all non-detected sequence variants removed, (2) with zeros as needed to fill out the 1,142-row ASV table, and (3) with zeros as needed to fill out the 3,342-row ASV table. The appropriate number of zeros to be included for each sample cannot be known, but the Shannon index estimated with all non-detected sequence variants removed is very likely underestimated. The results (**Figure 2**) show that the Shannon index can be quite precisely estimated (narrow error bars) with library sizes nearing 30,000 sequences but that the number of zeros included in the analysis can still have a substantial effect on accurate estimation of the Shannon index of the source (results, though precise, vary widely with the number of zeros included in the data). It is thus concluded that it is not statistically possible to estimate the Shannon index of the source (even if all the assumptions are met that enable use of the multinomial relative abundance model) unless the number of unique variants

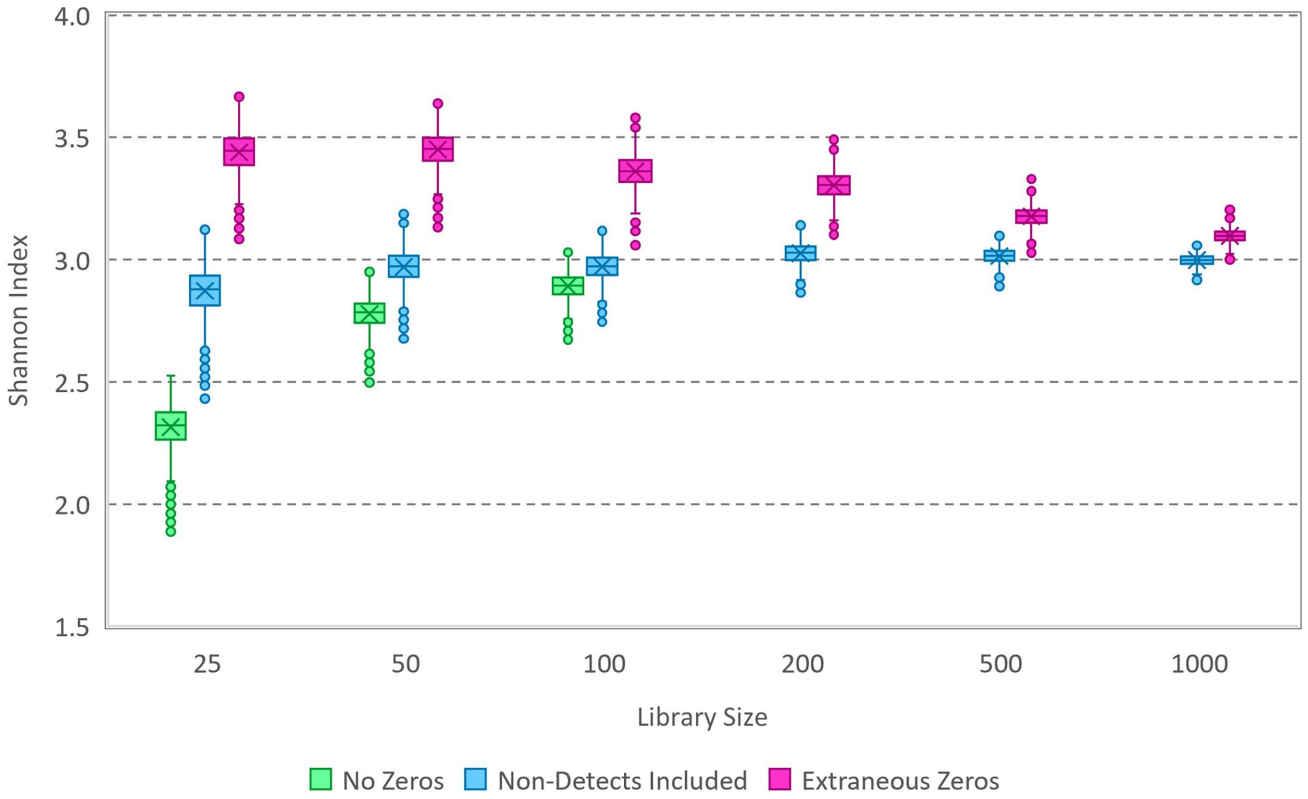

FIGURE 1 | Box and whisker plot of Markov chain Monte Carlo (MCMC) samples from posterior distributions of the Shannon index based on analysis of simulated data (**Supplementary Data Sheet 2**). Data with various library sizes (**Supplementary Table S1**) were analyzed in each of three ways: with zeros excluded (not applicable in some cases), with zeros included for non-detected variants, and with extraneous zeros corresponding to variants that do not exist in the source. The true Shannon index of the source from which the data were simulated is 3.03.

present in the source is precisely known *a priori*. Accordingly, the following section elaborates upon evaluation of sample-level diversity accounting for varying library sizes and the mechanistic process by which data are obtained.

DIVERSITY ANALYSIS IN ABSENCE OF A MODEL TO INFER SOURCE DIVERSITY

Recognizing that amplicon sequencing of a sample provides only partial and indirect representation of the diversity in the source (specifically partial representation of post-amplification diversity) and that statistical inference about source diversity is compromised by clustering, amplification, and not knowing how many zeros should be included in the data, the question of how to perform diversity analysis remains. The approach should acknowledge the random nature of amplicon sequencing data, reflect the importance of the library size in progressively revealing information about diversity, avoid normalization that distorts the proportional composition of samples, and provide some measure of uncertainty or error. Inference about source diversity is the ideal, but it is not possible with a multinomial relative abundance model unless the number of unique variants in the source is precisely known, and there are many types of error in amplicon sequencing that are likely to invalidate this foundational model as discussed above. Rarefying repeatedly, a subsampling process to normalize library sizes among samples that is performed many times in order to characterize the variability introduced by rarefying (Cameron et al., 2021), satisfies these goals. When a sample is rarefied repeatedly down to a smaller library size (using sampling without replacement), it describes what data might have been obtained if only the smaller library size of sequence variants had been observed. This can then be propagated to develop a range of values of an alpha diversity metric or a cluster of points on a beta diversity ordination plot to graphically display the variability introduced by rarefying to a normalized library size. It also does not throw out valid sequences because all sequences are represented with a sufficiently large number of repetitions. A value of the sample Shannon index may then be computed for each of the repetitions to quantify the diversity in samples of a particular library size.

Figure 3 illustrates the relationship between repeatedly rarefying to smaller library sizes and statistical inference about the source from which the sample was taken. Rarefying adds random variability by subsampling without replacement while statistical inference includes parametric uncertainty that is often ignored in contemporary diversity analyses. Because the extent to which

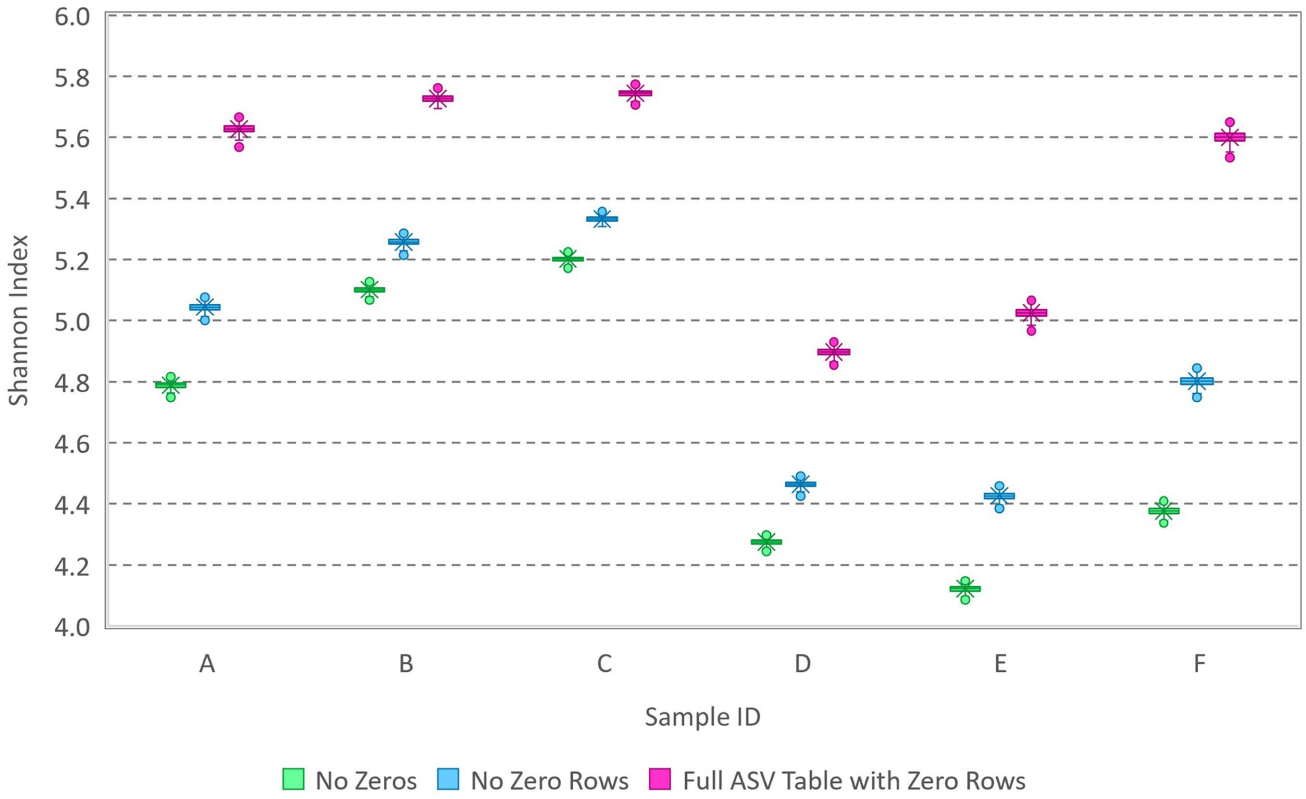

FIGURE 2 | Box and whisker plot of MCMC samples from posterior distributions of the Shannon index based on analysis of environmental amplicon sequencing data (**Supplementary Data Sheet 2**). Data with various library sizes between 10,000 and 30,000 were analyzed three ways: with zeros excluded, with zeros included in a 1,142-row ASV table (no zero rows), and with additional zeros from the full 3,342-row ASV table including variants with rows of zeros (detected in other samples from the same study area).

diversity is exhibited by a sample depends on the library size, such sample-level analysis must be performed at the same level (analogous to converting 1 mm, 1 cm, and 1 km to a common unit before comparing numerical values) and any observations obtained about patterns in sample-level diversity are conditional on the shared library size at which the analysis was performed. On the other hand, current methods (including rarefying once), distort the data to facilitate use of compositional analysis methods that presume the transformed data are a perfect representation of the microbial composition in the source; it is important to recognize that the detected library is only a random sample that is imperfectly representative of source diversity.

Cameron et al. (2021) addressed issues such as choice of normalized library size and sampling with or without replacement and completed both alpha and beta diversity analyses using repeated rarefaction and environmental data, but it did not include simulation experiments. Accordingly, a simulation experiment was performed using the hypothetical population-based on Scrabble™ and samples with varying library sizes (see R code in **Supplementary Material**) to explore rarefying repeatedly and plug-in estimation of the Shannon index (**Figure 4**). A thousand simulated datasets with a library size of 25 yielded Shannon index values between 1.86 and 2.87 (with a mean of 2.49), illustrating that the source diversity (with a Shannon index of 3.03) is only partially exhibited by a sample with a library size of 25. Five samples were generated with library sizes of 50, 100, 200, 500, and 1,000, and corresponding Shannon index values are shown in red (deteriorating markedly at library sizes of 100 or less). Each sample was then rarefied repeatedly (1,000 times) to a library size of 25, resulting in the box and whisker plots of the calculated Shannon index values. Although samples with larger library sizes exhibit more diversity, samples repeatedly rarefied down to the minimum library size of 25 exhibit very comparable diversity. The Shannon index at a library size of 25 is similar for all samples, as it should be given that they were generated from the same population. If rarefying had been completed only once without quantification of the error introduced, it may erroneously have been concluded that the samples exhibited different Shannon index values.

DISCUSSION

Diversity analysis of amplicon sequencing data has grown rapidly, adopting tools from other disciplines but largely differing from the statistical approaches applied to classical microbiology data. Most analyses feature a deterministic set of procedures to transform the data from each sample and yield a single value of an alpha diversity metric or a single

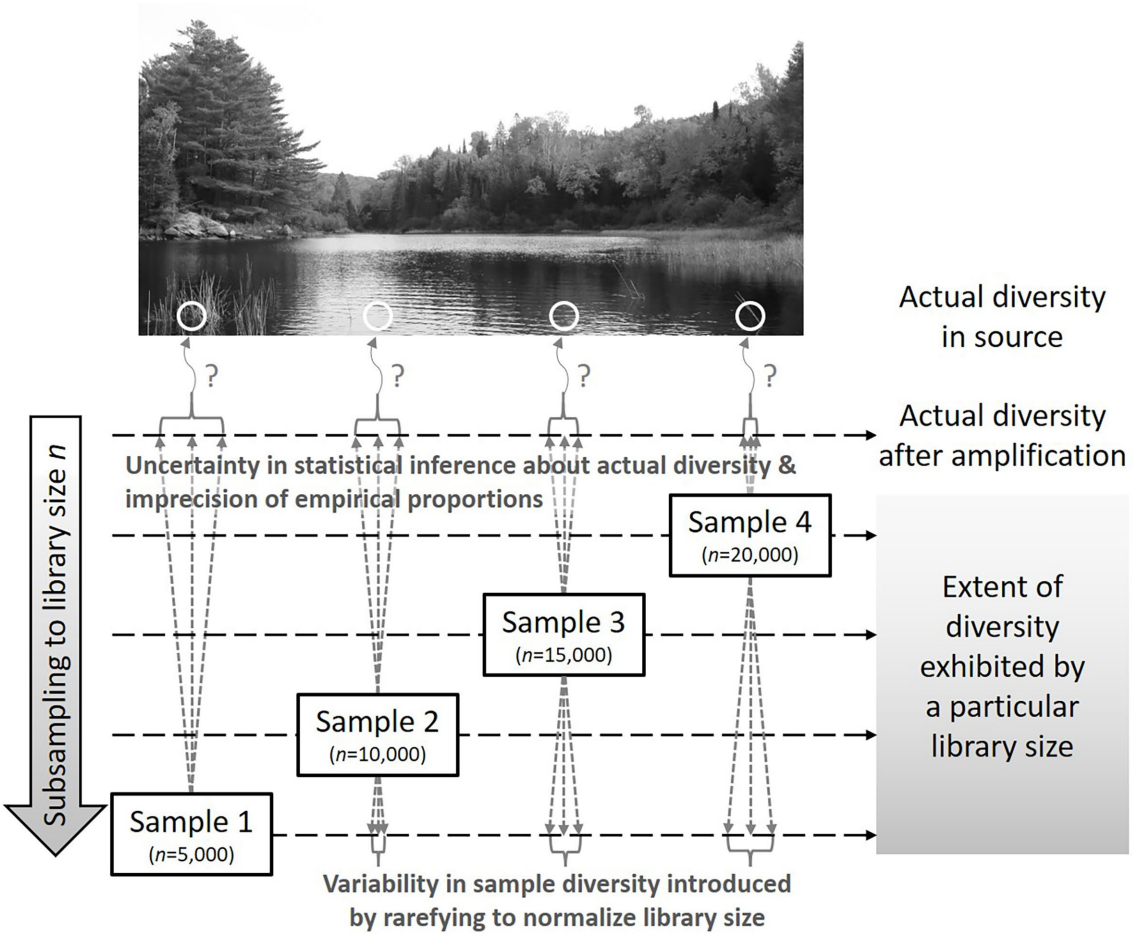

FIGURE 3 | Representation of how library size and diversity quantified therefrom relate to uncertainty in statistical inference about source diversity and variability introduced by repeatedly rarefying to the smallest obtained library size. In this case, rarefying repeatedly evaluates the extent of the diversity (after amplification) exhibited if a library size of only $n = 5{,}000$ had been obtained from each sample.

point on an ordination plot. Such procedures should not be viewed as statistical analysis because observed sequence count data are not a population (i.e., perfect measurements of the proportional composition of the community in the source); they are a random sample representing only a portion of that population. Recognizing that the data are random and that the goal is to understand the alpha and beta diversity of the sources from which samples were collected, it is important to describe and explore the error mechanisms leading to variability in the data and uncertainty in estimated diversity. Additionally, graphical displays of results should reflect uncertainty, such as by presenting error bars around alpha diversity values.

This study provides a step toward such methods by describing mechanistic random errors and their potential effects, proposing a probabilistic model and listing the assumptions that facilitate its use, discussing various types of zeros that may appear (or fail to) in ASV tables, and performing illustrative analyses using simulated and environmental data. Several sources of random error were found to invalidate the multinomial relative abundance model that is foundational to probabilistic modeling of compositional sequence count data, notably including clustering of microorganisms in the source and amplification of genes in this sequencing technology. Nonetheless, it is a good starting point for inference of the alpha diversity of the source and quantifying uncertainty therein. Future simulation studies could explore the effect of non-random microorganism dispersion, sample volume (relative to a hypothetical *representative elementary volume* of the source), differential analytical recovery in sample processing, amplification errors, and sequencing errors on diversity analysis more thoroughly and evaluate the potential for current normalization and point estimation approaches to misrepresent diversity.

This study also presents a simple Bayesian approach to drawing inference about diversity in the sources from which samples were collected (rather than just diversity in the sample or some transformation of it). Even under idealized circumstances in which the multinomial relative abundance model is valid, it was unfortunately found to be biased unless the number of unique variants present in the source was known *a priori*. This may have implications on analysis of

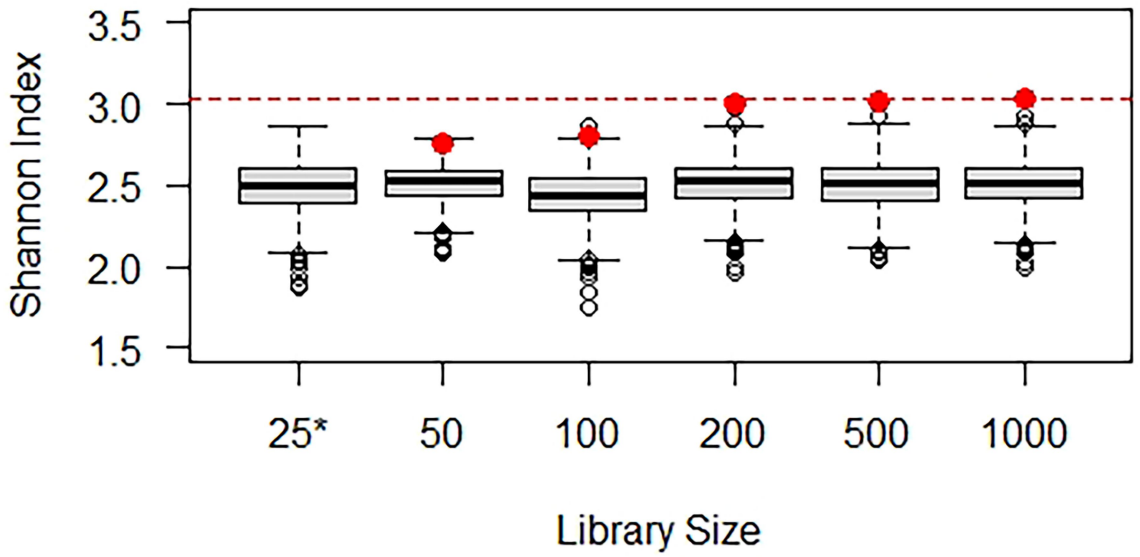

FIGURE 4 | Demonstration of normalization by rarefying repeatedly using simulated data. The box and whisker plot for the library size of 25 (*) illustrates how the Shannon index varies among simulated samples and is consistently below the actual Shannon index of 3.03 (red line). The Shannon index calculated from the samples with larger library sizes (red dots) deteriorates at small library sizes. The box and whisker plots for these library sizes illustrate what Shannon index might have been calculated if only a library size of 25 had been obtained (rarefying 1,000 times to this level). In all cases, a Shannon index of about 2.5 is expected with a library size of 25.

any type of multinomial data, beyond microbiome data, in which the number of possible outcomes (or the number of outcomes with zero observations that should be included in the analysis) is unknown. It is plausible that a probabilistic model could be developed to account for errors that invalidate the multinomial model, though this would require many assumptions that would be difficult to validate and that could substantially bias inferences if the assumptions are incorrect. In summary, probabilistic modeling should be used to draw inferences about source diversity and quantify uncertainty therein, but the simple multinomial model is invalidated by some types of error that are inherent to the method and inference is not possible even with the multinomial model unless the practically unknowable number of unique variants in the source is known.

For lack of a reliable and readily available probabilistic approach to draw inferences about source diversity, an approach to evaluate and contrast sample-level diversity at a particular library size is needed. Rarefying only once manipulates the data in a way that adds variability and discards data (McMurdie and Holmes, 2014), and (like other transformations proposed to normalize data) the manipulated data are generally only used to obtain a plug-in estimate of diversity. Rarefying repeatedly, on the other hand, allows comparison of sample-level diversity estimates conditional on a library size that is common among all analyzed samples, does not discard data, and characterizes variability in what the diversity measure might have been if only the smaller library size had been observed. This approach is by no means statistically ideal, but it may be a distant second best relative to the Bayesian approach (or analogous frequentist approaches based on the likelihood function) presented in this study that cannot practically be applied in an unbiased way in many scenarios, especially due to the unknowable number of unique variants that are actually present in the source and complex error structures inherent to amplicon sequencing.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the **Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PS and EC developed the theoretical model with support from ME and KM. PS conceived the statistical analysis approach, carried out or directed the simulation experiments and data analysis, and developed the visuals with support from ME and EC. PS wrote the manuscript with support from ME, EC, and KM. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge the support of NSERC (Natural Science and Engineering Research Council of Canada), specifically through the *for*Water NSERC Network for Forested Drinking Water Source Protection Technologies (NETGP-494312-16), and Alberta Innovates

(3360-E086). This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program (ME; Canada Research Chair in Water Science, Technology & Policy).

ACKNOWLEDGMENTS

We are grateful for discussion of this work with Mary E. Thompson (Department of Statistics and Actuarial Science, University of Waterloo) and assistance with performing and graphing analyses from Rachel H. M. Ruffo.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.728146/full#supplementary-material

REFERENCES

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

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

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics Inform.* 17:e6. doi: 10.5808/GI.2019.17.1.e6

Cameron, E. S., Schmidt, P. J., Tremblay, B. J. M., Emelko, M. B., and Müller, K. M. (2021). Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Sci. Rep.* 11:22302. doi: 10.1038/s41598-021-01636-1

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Chik, A. H. S., Schmidt, P. J., and Emelko, M. B. (2018). Learning something from nothing: the critical importance of rethinking microbial non-detects. *Front. Microbiol.* 9:2304. doi: 10.3389/fmicb.2018.02304

Eisenhart, C., and Wilson, P. W. (1943). Statistical methods and control in bacteriology. *Bacteriol. Rev.* 7, 57–137. doi: 10.1128/br.7.2.57-137.1943

Emelko, M. B., Schmidt, P. J., and Reilly, P. M. (2010). Particle and microorganism enumeration data: enabling quantitative rigor and judicious interpretation. *Environ. Sci. Technol.* 44, 1720–1727. doi: 10.1021/es902382a

Fisher, R. A., Thornton, H. G., and Mackenzie, W. A. (1922). The accuracy of the plating method of estimating the density of bacterial populations. *Ann. Appl. Biol.* 9, 325–359. doi: 10.1111/j.1744-7348.1922.tb05962.x

Freedman, D. A., Pisani, R., and Purves, R. A. (1998). *Statistics. 3rd Edn*. New York: W. W. Norton, Inc.

Fricker, A. M., Podlesny, D., and Fricke, W. F. (2019). What is new and relevant for sequencing-based microbiome research? A mini-review. *J. Adv. Res.* 19, 105–112. doi: 10.1016/j.jare.2019.03.006

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Helsel, D. (2010). Much ado about next to nothing: incorporating nondetects in science. *Ann. Occup. Hyg.* 54, 257–262. doi: 10.1093/annhyg/mep092

Huggett, J. F., O'Grady, J., and Bustin, S. (2015). qPCR, dPCR, NGS: a journey. *Biomol. Detect. Quantif.* 3, A1–A5. doi: 10.1016/j.bdq.2015.01.001

Hughes, J. B., and Hellmann, J. J. (2005). The application of rarefaction techniques to molecular inventories of microbial diversity. *Meth. Enzymol.* 397, 292–308. doi: 10.1016/S0076-6879(05)97017-1

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* 8:2114. doi: 10.3389/fmicb.2017.02114

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

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26:27663. doi: 10.3402/mehd.v26.27663

McCrady, M. H. (1915). The numerical interpretation of fermentation-tube results. *J. Infect. Dis.* 17, 183–212. doi: 10.1093/infdis/17.1.183

McGregor, K., Labbe, A., Greenwood, C. M. T., Parsons, T., and Quince, C. (2020). Microbial community modelling and diversity estimation using the hierarchical pitman-yor process. bioRxiv. doi: 10.1101/2020.10.24.353599

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. (2019). Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.* 10, 389–400. doi: 10.1111/2041-210X.13115

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *Elife* 8:e46923. doi: 10.7554/eLife.46923.001

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Nahrstedt, A., and Gimbel, R. (1996). A statistical method for determining the reliability of the analytical results in the detection of *Cryptosporidium* and *Giardia* in water. *J. Water Supply: Res. Technol.* 45, 101–111.

Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *Am. Nat.* 102, 243–282. doi: 10.1086/282541

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43:e37. doi: 10.1093/nar/gku1341

Schmidt, P. J., Emelko, M. B., and Thompson, M. E. (2014). Variance decomposition: a tool enabling strategic improvement of the precision of analytical recovery and concentration estimates associated with microorganism enumeration methods. *Water Res.* 55, 203–214. doi: 10.1016/j.watres.2014.02.015

Schmidt, P. J., Emelko, M. B., and Thompson, M. E. (2020). Recognizing structural nonidentifiability: when experiments do not provide information about important parameters and misleading models can still have great fit. *Risk Anal.* 40, 352–369. doi: 10.1111/risa.13386

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

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

Student (1907). On the error of counting with a haemacytometer. *Biometrika* 5, 351–360. doi: 10.2307/2331633

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., AL-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4:62. doi: 10.1186/s40168-016-0208-8

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

United States Environmental Protection Agency (2005). Technical Report EPA 815-R-05-002. *Cryptosporidium* and *Giardia* in Water by Filtration/IMS/FA.

Washington, H. G. (1984). Diversity, biotic and similarity indices: a review with special relevance to aquatic ecosystems. *Water Res.* 18, 653–694. doi: 10.1016/0043-1354(84)90164-7

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* 10:2407. doi: 10.3389/fmicb.2019.02407

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.