forWater
NSERC Network for Forested Drinking Water Source Protection Technologies

UNIVERSITY OF WATERLOO

forWater Research Snapshot

AUTHORS: Ellen Cameron (PhD Graduate, UW), Philip Schmidt (UW), Monica Emelko (UW), Kirsten Müller (UW)

SUBJECT: **Two publications: Modelling the variability of next-generation sequencing (NGS) data used in microbial diversity analysis & Enhancing diversity analysis by repeatedly rarefying amplicon data &**

Research Summary – Statistical modelling of amplicon sequencing and microbial diversity analysis

Widespread adoption of well-established next generation sequencing (NGS) technologies is increasingly advocated, including in the water industry. NGS involves parallel sequencing of multiple small fragments of DNA to determine the unique order of its building blocks in target regions or entire genomes; parallel sequencing dramatically reduces analysis time.

Amplicon sequencing is a highly targeted approach used to analyze genetic variation in specific genomic regions such as the 16S rRNA gene to study bacterial phylogeny (i.e., the history of the evolution and relationships among broad groups of organisms) and taxonomy (i.e., classification). Thus, amplicon sequencing and other NGS approaches are essential tools for identifying and managing algae-associated risks (e.g., formation potential of toxins as well as taste and odour) and other microbially mediated processes (e.g., nutrient dynamics, contaminant transformation in drinking water source watersheds and storage reservoirs). Amplicon sequencing yields large libraries of sequences—there are many approaches for interpreting the resulting data and some are more biased than others. Despite the variety of available approaches, foundational concepts of microbiology (such as microorganisms and their genes being discrete objects thereby making it impossible to have 0.5 microorganisms) and established knowledge surrounding their enumeration are not reflected in many diversity analysis approaches.

Here, the unavoidable random errors affecting the relative abundance of microbial sequences accounting for 1) variability introduced by collection and handling of environmental samples, 2) losses during sample processing, 3) amplification, and 4) sequencing were investigated. Attention was given to the handling of zeros (i.e., sequences not observed in a sample) to emphasize their importance in *unbiased* diversity analysis. It was shown that unbiased estimation of source diversity in natural systems is impeded because the exact number of unique variants in the source is never known. Sample-level diversity analysis using repeated rarefaction that yields bands/regions of values that characterize variation of diversity was recommended for interpretation of amplicon sequencing data.

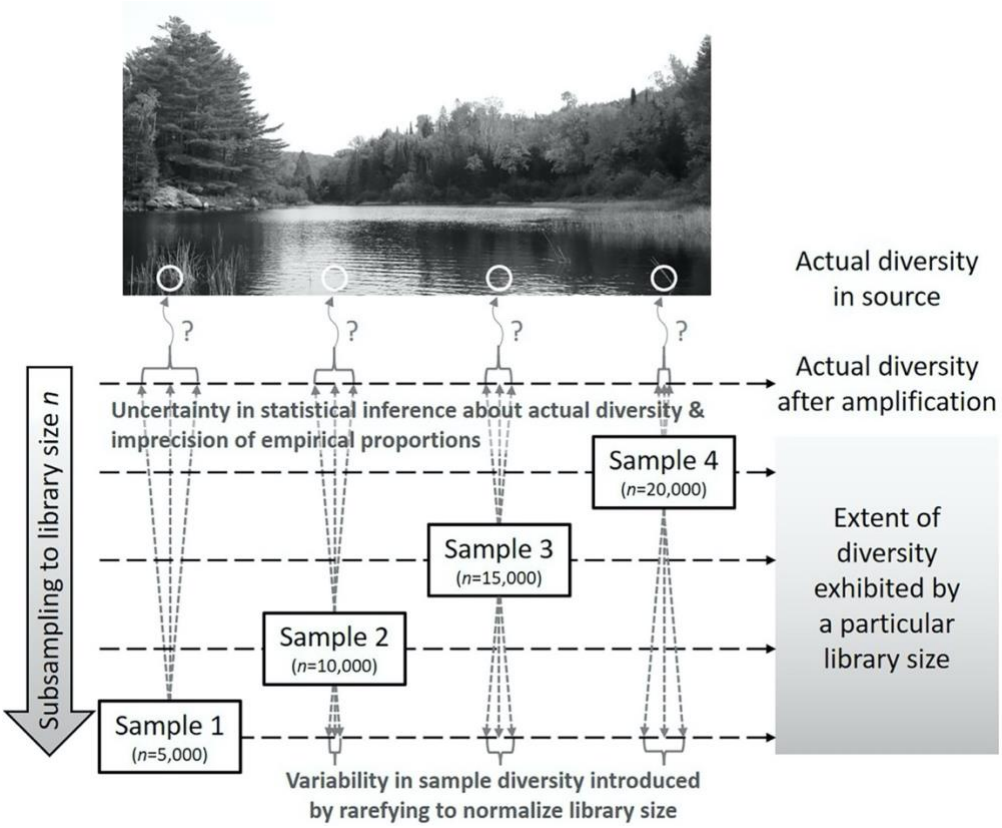
Research results

- Recognizing that obtained sequence data are a random subset of sequences obtained following a complex laboratory method 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 1) the error mechanisms leading to variability in the data and 2) uncertainty in estimated diversity
- There is substantial knowledge in classical quantitative microbiology methods that can be informative about the nature of amplicon sequencing data used in microbial diversity analysis and quantification of uncertainty in results
- Zeros in sequencing tables can reflect true absence of a particular sequence in the source or non-detected sequences, but there can also be non-detected sequences for which no zero appears in the sequencing tables
- Unbiased estimation of alpha diversity (e.g., the Shannon index) is impeded by zeros unless the number of unique variants in the source is precisely known
- Rarefying repeatedly does not discard data and characterizes the sample-level diversity that might have been observed if only the smaller library size common among all analyzed samples had been observed

Key messages

- As is increasingly recognized elsewhere in quantitative microbiology, consideration of the mechanisms leading to variation in data, modelling that reflects the discrete nature of microorganisms and their genes, and appropriate handling of zeros are essential to unbiased data analysis
- There are many types of random error in sample collection, handling, processing, amplification, and sequencing that can affect how representative sequence libraries are of microbial diversity in the source, and continued research in this area is needed to extract reliable information from obtained sequencing data
- Zeros that do not even appear in sequence tables are newly recognized to be critically important for unbiased diversity analysis
- Rarefying repeatedly provides appropriate representation of the level of diversity revealed at a particular library size at which all samples are compared

Figure 1



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.

Reference: Schmidt PJ, Cameron ES, Müller KM, Emelko MB, 2022. Ensuring that Fundamentals of Quantitative Microbiology are Reflected in Microbial Diversity Analyses based on Next-generation Sequencing. *Frontiers in Microbiology*, 13:728146. https://doi.org/10.3389/fmicb.2022.728146

Research Summary – Repeated rarefying to assess microbial diversity without discarding valid data

Amplicon sequencing has revolutionized our ability to study DNA collected from environmental samples. This can be especially useful in evaluating ecosystem change and signaling potential water quality concerns. Amplicon sequencing data consist of discrete counts of sequence reads, the sum of which is the library size. Library sizes typically vary substantially and normalization is used so that measures of diversity are not affected by varying library sizes. While the widely used approach of subsampling randomly from observed libraries—rarefaction—is simple, it has been criticized for discarding valid data. Critically, however, many other types of normalization transform data and involve manipulation (e.g., omission, arbitrary modification) of zeros.

Here, *repeated* rarefaction to enhance assessment of similarities or differences in microbial community species diversity between samples was investigated. While single rarefaction excludes some data, repetition of this process ensures representation of all data *without manipulation*. Rather than representing diversity as a single numerical value or point on a plot, rarefying repeatedly yields bands/regions of values that characterize variation of diversity within or between samples at a particular library size.

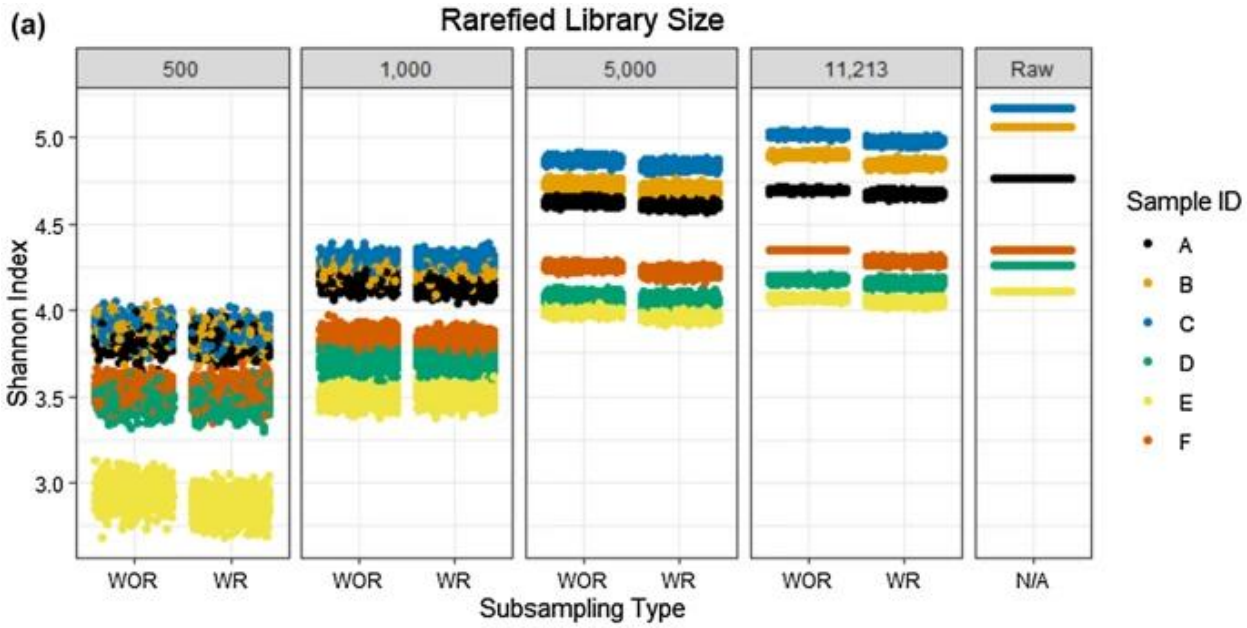
Research results

- Repeated rarefaction enables proportionate representation of all observed sequences with characterization of the random variation introduced to diversity analyses by rarefying to a normalized library size; however, this describes diversity observed in libraries of a particular size rather than inference about source diversity
- This process reflects which data might have been obtained if a sample's library size had been smaller and allows graphical representation of the effects of library size normalization on diversity analysis results
- Rarefying repeatedly produces bands of alpha diversity metrics (e.g., the Shannon index) and regions on beta diversity ordination plots (e.g., using Bray-Curtis distances) that show the variation introduced by rarefying
- Statistical analysis of amplicon sequencing is still in the early stages of development, and there is particular need for ongoing research in diversity analysis due to the need to normalize library sizes of samples
- Repeated rarefaction can be applied with small, normalized library sizes if necessary to avoid discarding samples, but the greater variation added by rarefying can impede observation of differences among samples

Key messages

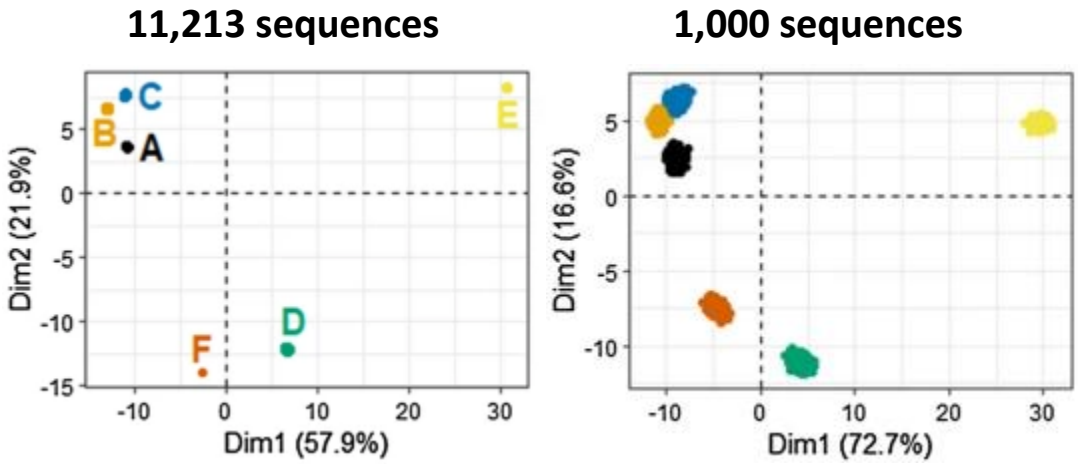
- When using NGS to understand water quality-associated changes in microbial communities, repeated rarefaction of data can help to describe community diversity present in different environmental samples as well as how normalization impacts the analysis
- The implementation of only a single iteration of rarefying is problematic due to the omission of valid data and occasional need to discard samples with small library sizes
- Use of larger normalized library sizes when rarefying minimizes the variation introduced by this process, but analysis at several normalized library sizes may be needed to be inclusive of samples with small libraries
- Further development of strategies (e.g., data handling, library size normalization for diversity analyses) is required for ensuring rigorous interpretation of amplicon sequencing data

Figure 2



Effect of chosen rarefied library size and sampling with or without replacement (WR and WOR, respectively) on the Shannon Diversity Index. Six microbial communities (A-F) were rarefied repeatedly at specific rarefied library sizes of 11,213 sequences (the smallest sample library size), 5000 sequences, 1000 sequences, and 500 sequences.

Figure 3



Variation in principal component analysis ordinations (using Bray–Curtis dissimilarity on Hellinger transformed rarefied libraries) of six microbial communities (A-F) repeatedly rarefied to varying library sizes.

Reference: Cameron ES, Schmidt PJ, Tremblay BJ-M, Emelko MB, Müller KM. 2021. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Scientific Reports*, 11:22302. https://doi.org/10.1038/s41598-021-01636-1

Contact Information

For more information on this research please contact:
Monica B. Emelko – mbemelko@uwaterloo.ca; Philip J. Schmidt - pj2schmidt@uwaterloo.ca