Waste not, want not: Revisiting the analysis that called rarefaction into question

Running title: Review of "Waste not, want not"

Patrick D. Schloss†

† To whom corresponsdence should be addressed:

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Research article

Abstract

Introduction

Since the development of sequencing technologies such as those provided by 454 and Illumina, microbiome researchers have struggled to produce a consistent number of sequences from each sample in a dataset. It is common to observe more than 10-fold variation in the number of sequences per sample [XXXXX]. Regardless of the source of this variation, researchers desire approaches to control for uneven sampling effort. Of course, this desire is not unique to microbiome research and is a challenge faced by all community ecologists. Common approaches to controlling uneven sampling efforts have included use of proportional abundance (i.e., relative abundance), normalization of counts, and rarefaction.

In 2014 Paul McMurdie and Susan Holmes published their "Wast not, want not: why rarefying microbiome data is inadmissible" (WNWN) in PLOS Computational Biology. This paper has had a significant impact on the approaches that microbiome researchers use to analyze 16S rRNA gene sequence data. According to Google Scholar, this paper has been cited more than 2,300 times as of January 2023. Anecdotely, I have received correspondence from researchers over the past 10 years asking how to address critiques from reviewers who criticize my correspondents' analysis for rarefying. I have also received these types of comments from reviewers, specifically in regards to a preprint that I posted in 202X in regards to my critique of the practice of removing rare taxa from analyses. In the process of preparing a manuscript investigating rarefaction and other approaches to control for uneven sequencing effort, I decided to reassess the WNWN study including their definitions, simulations, and analyses.

Re-running the R code that the authors published as Protocol S1, I noticed that the step that purported to rarefy the data only performed one subsampling of the data (Lines 404 through 416 of `simulation-cluster-accuracy/simulation-cluster-accuracy-server.Rmd`). This caused me to re-inspect how McMurdie and Holmes defined "rarefying" in the following quoted text from their paper:

> Instead, microbiome analysis workflows often begin with an ad hoc library size normalization by random subsampling without replacement, or so-called rarefying [17]–[19]. There is confusion in the literature regarding terminology, and sometimes this normalization approach is conflated with a non-parametric resampling technique — called rarefaction [20], or individual-based taxon re-sampling curves [21] — that can be justified for coverage analysis or species richness estimation in some settings [21], though in other settings it can perform worse than parametric methods [22]. Here we emphasize the distinction between taxon re-sampling curves and normalization by strictly adhering to the terms rarefying or rarefied counts when referring to the normalization procedure, respecting the original definition for rarefaction. Rarefying is most often defined by

the following steps [18].

1. Select a minimum library size, $N_{L,m}$. This has also been called the rarefaction level [17], though we will not use the term here.
2. Discard libraries (microbiome samples) that have fewer reads than $N_{L,m}$.
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,m}$.

Often $N_{L,m}$ is chosen to be equal to the size of the smallest library that is not considered defective, and the process of identifying defective samples comes with a risk of subjectivity and bias. In many cases researchers have also failed to repeat the random subsampling step (3) or record the pseudorandom number generation seed/process — both of which are essential for reproducibility.

It is unfortunate that McMurdie and Holmes used the term rarefying here and throughout their manuscript. The authors were correct to state that the distinction between “rarefying” and “rarefaction” is confusing and leads to their conflation. In my experience, subsequent researchers have conflated the results of this study of the effects of rarefying data with rarefaction. Adding to the confusion is that the papers cited in their first sentence either do not use the words “rarefy” or “rarefying” or use them interchangably with “rarefaction”. In hindsight, as shown in the quoted text, McMurdie and Holmes do emphasize the distrinction between rarefying and rarefaction. However, because they seem to have coined a new meaning for rarefying, they seem to have only added to the confusion by using the generally used verb form of rarefaction. Further confusion comes from the author’s admonition in the final sentence that some researhcers have failed to repeat the subsampling step. To most scientists, repeating the subsampling step is rarefaction. My preference is to use subsampling as the term describing the process they refer to as rarefying. In other words rarefaction with a single randomization.

To provide a more clear definition of rarefaction, I propose the following:

1. Select a minimum library size, $N_{L,m}$.
2. Discard samples that have fewer reads than $N_{L,m}$.
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,m}$.
4. Compute the desired metric (e.g., richness, Shannon diversity, Bray-Curtis distances) using the subsampled data
5. Repeat steps 3 and 4 a large number of iterations (e.g, 100 or 1,000)
6. Compute summary statistics using values generated from the subsampled data

With this approach, rarefaction can be performed using any alpha or beta diversity metric. Furthermore, the procedure could also be used for hypothesis tests of differential abundance; however, thought would need to be given to how to synthesize the results of these tests across a large number of replications. Researchers are encouraged to report the minimum library size as well as the number of iterations used in their analysis. It is important to note that this procedure generates substantially different results to those obtained without accounting for uneven sampling effort or using relative abundances, normalized counts, compositional data transformatins, and variance stabilization procedures. I have explored the differences in results obtained using the diversity of approaches for controlling for uneven sampling effort; rarefaction, as described here, outperforms the other approaches [XXXXX].

Additional issues * Removed rare/patchy taxa from dataset after subsampling the data - * Removed samples only from subsampled datasets, which they admitted would detrimentally affect performance * Model for defining differential abundance is weird

In this review of WNWN will reassess three things... * Toy example * Cluster analysis * Differential abundance

Results

Toy example

Cluster analysis

Differential abundance

Discussion

- Gratidue that code was published with paper. This made it straight forward to notice that only a single subsampling step was performed for each random seed and that only 3 random seeds were used.

Acknowledgements

References

Figures