Waste not, want not: Revisiting the analysis that called the practice of rarefying microbiome data 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 highly parallelized sequencing technologies that enable microbiome researchers to distribute sequences across multiple samples in a single sequencing run, 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]. Researchers desire strategies to limit uneven sampling effort and need methods to control for he uneveness in their analyses. Of course, uneven sampling is not unique to microbiome research and is a challenge faced by all community ecologists. Common approaches to controlling uneven sampling efforts include use of proportional abundance (i.e., relative abundance), normalization of counts, parameter estimation, and rarefaction.

In 2014 Paul McMurdie and Susan Holmes published "Waste not, want not: why rarefying microbiome data is inadmissible" (WNWN) in PLOS Computational Biology [XXXXX]. 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 2,400 times as of April 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 using rarefaction (e.g., see this Twitter thread). I have also received similar comments from reviewers regarding my own work. Most recently, I received the critique for a preprint that I posted in 20XX that analyzes the practice of removing rare taxa from microbiome analyses [XXXXX]. In the process of responding to these reviewers' comments and 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.

Confusion regarding what is meant by "rarefying" and "rarefaction"

As I attempted to reproduce the results of WNWN, I noticed that the step that rarefied the data only performed one subsampling of the data. 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 resampling 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 was 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. Adding to the confusion is that the papers cited in the first sentence of this quote (i.e., their references [17]–[19]) either do not use the words “rarefy” or “rarefying” or use them interchangably with “rarefaction”. In my experience, subsequent researchers have conflated the results of this study of the effects of rarefying data with rarefaction of data. As an example, Willis (XXXXX) describes problems with rarefaction rather than rarefying data when citing WNWN in her paper proposing alternatives to rarefaction for use with alpha diversity data:

> Unfortunately, rarefaction is neither justifiable nor necessary, a view framed statistically by McMurdie and Holmes (2014) in the context of comparison of relative abundances.

In hindsight, as shown in the quoted text from WNWN, McMurdie and Holmes did emphasize the distinction between rarefying and rarefaction. However, because they seem to have coined a new meaning for rarefying, they 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 researchers 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, subsampling

is rarefaction, but with a single randomization. To minimize confusion, I will use subsamping in place of rarefying.

I propose the following definition of rarefaction:

1. Select a minimum library size, $N_{L,m}$. Researchers are encouraged to report the value of $N_{L,m}$.
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). Researchers are encouraged to report the number of iterations.
6. Compute summary statistics using values generated from the subsampled data

This definition aligns well with how rarefaction was originally defined for comparing richness (i.e., the number of taxa in a community) across communities when communities are sampled to different depths [XXXXXXXXX]. With this more general approach to rarefaction, rarefaction can be performed using any alpha or beta diversity metric. This strategy has been widely used by my research group and others and is available in the mothur software package using commands such as `summary.single`, `rarefaction.single`, `dist.shared`. The procedure outlined above 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.

Description of "Simulation A" from WNWN

McMurdie and Holmes analyzed the effect of rarefying and other approaches on clustering accuracy using what they called "Simulation A" in their Figure 2A and elsewhere in their paper. In Simulation A, they investigated the ability to correctly assign simulated microbiome samples to one of two clusters representing two simulated treatment groups. Thankfully, the R code for Simulation A was provided by the authors in the `simulation-cluster-accuracy/simulation-cluster-accuracy-server.Rmd` R-flavored markdown file that was published as Protocol S1 in the original paper. I will outline the simulation strategy below and will reference line numbers from their R-flavored markdown file with "L" as a prefix.

For the WNWN cluster analysis, sampling distributions for the two treatment groups were generated using human fecal and ocean data originally take from the GlobalPaterns dataset (L129) [XXXXX]. To generate

a fecal and ocean distribution, the authors included any operational taxonomic unit (OTU) that appeared in more than one of the 4 fecal and 3 ocean samples (L60 and L137). The OTUs were sorted by how many of the 7 samples the OTUs were observed in followed by their total abundance across all 7 samples (L139). From this sorted list they identified the identifiers of the first 2000 OTUs (L66). Returning to the 7 samples they selected the data for the corresponding 2000 OTUs and pooled the OTU abundances of the fecal and ocean samples separately to create two sampling distributions (L144, L159-160, L197-198). Next, the fecal and ocean distributions were mixed in 8 different fractions to generate two community types that differend by varying effect sizes (i.e., 1, 1.15, 1.25, 1.5, 1.75, 2, 2.5, and 3.5; L170-195, L220); an effect size of 1 generated a null model with no difference between the treatment groups. To simulate the variation in sequencing depth across the 80 samples, they normalized he number of sequences from each of the 26 samples in the GlobalPatterns dataset so that the median number of sequences ($\tilde{N}_L$) for the GlobalPatterns had 1,000, 2,000, 5,000, or 10,000 sequences (L324-325). They then randomly sampled the 26 normalized sequencing depths to generate 80 sampling depths. From each community type, they simulated 40 samples by sampling to the desired number of sequences (L73, L230-233 and L326-327). Each simulation condition was repeated 5 times (L85). This resulted in 160 simulations (8 effect sizes x 4 median sampling depths x 5 replicates = 160 simulations). Finally, they removed rare and low prevalence OTUs in two steps. First, they removed any OTUs whose total abundance was less than 3 across all 80 samples and that did not appear in at least 3 samples (L368-386). Second, they removed any OTUs that did not have more than 1 sequence in more than 5% of the 80 samples (i.e., 4 samples) and that did not have a total abundance across the 80 samples greater than one half of the number of samples in each community type (i.e., 20) (L523-538, L551).

After generating the OTU counts for the 160 simulated communities, the authors applied several normalization methods, distance calculations, and clustering algorithms to the data. To normalize the OTU data the original analysis either applied no normalization procedure, calculated OTU relative abundances, subsampled the data, performed variance stabilization using the DESeq R package, or performed Upper-Quartile log-fold change normalization using the edgeR R package. In subsampling the data, the authors either included all of the samples or removed samples whose sequencing depth fell below the 5, 10, 15, 20, 25, and 40 percentile across all 80 samples. Subsampled data were then used as input to calculate distances between samples using Bray-Curtis, Euclidean, Unweighted UniFrac, and Weighted Unifrac distances as implemented in the Phyloseq R package, Poisson distance as implemented in the PoiClaClu R package, and top-mean squared difference as implemented in the edgeR R package. Un-normalized data were used to calculate Bray-Curtis, Euclidean, Poisson, and Weighted UniFrac distances. Relative abundance data

were used to calculate Bray-Curtis, Unweigthed UniFrac, and Weighted UniFrac distances. DESeq variance stabilization normalized data were used to calculate Bray-Curtis, Euclidean, and Weighted Unifrac distances. The Upper-Quartile log-fold change normalized data were only used to calculate top-mean squared difference distances. The resulting distance matrices were used to cluster the 80 samples into one of two clusters using partitioning around the medoid (PAM), K-means clustering, and hierarchical clustering. Although data for all three methods were presented in the supplementary Protocol S1, only the PAM data are presented in the main manuscript. The accuracy of the clustering assignments was quantified as the fraction of the 80 samples that were assigned to the correct cluster. Because some of the subsampling conditions removed samples those were counted as mis-clustered samples yielding minimum accuracies below 50%.

Critique of the original simulation design

Although all simulations represent an artificial representation of reality and can be critiqued, eleven elements of the design of Simulation A warrant further review.

1. Simulated conditions were only replicated 5 times each
2. The average sizes of the libraries were small by modern standards
3. DESeq-based variance stabilization was used with distance calculation methods that are sensitive to negative values
4. A single subsampling of each dataset was evaluated rather than using rarefaction, which likely resulted in noisier data
5. Results using PAM clustering were not directly compared to those of K-means and hierarchical clustering
6. Subsampling removed the smallest 15% of the samples, which penalized accuracy values by 15 percentage points
7. The distribution of library sizes was not typical of those commonly seen in microbiome analyses
8. A filtering step was applied to remove rare taxa from the simulated datasets
9. No accounting for difference in performance when library sizes are confounded with treatment group
10. Clustering accuracy was used rather than direct comparisons of beta diversity
11. There was no consideration of effects of normalization methods on alpha diversity metrics

These points will serve as an outline for the Results section. After replicating the original simulations, these points will be evaluated to reassess whether subsampling or rarefaction are “inadmissible”.

Results

Replication of WNWN simulations and results

Before assessing the impact of the points I critiqued above, I attempted to replicate the results shown in Figures 4 and 5 of the original paper using the authors' original code. I created a Conda environment that used the R version and package versions that were as close as possible to those used in the original paper. Because of the slight differences in packages, it was necessary to apply several patches to the original R-flavored markdown file to render document. I was able to generate a figure similar to that presented as Figures 4 and 5 of the original paper. My results are shown in Figures S2 (**norarefy-source/simulation-cluster-accuracy/Figure_3.pdf**) and S3 (**norarefy-source/simulation-cluster-accuracy/Figure_4.pdf**) of this paper, respectively. The differences in results are likely due to differences in software versions and operating systems. It is also worth noting that the published versions of the two figures differ from those included in Protocol S1 within the rendered html file (`simulation-cluster-accuracy/simulation-cluster-accuracy-server.html`) and that the figure numbers are one higher in the paper than those generated by the R-flavored markdown file (i.e., Protocol S1 has Figures 3 and 4 corresponding to the published Figures 4 and 5). Regardless of the differences, my results were qualitatively similar to the originals.

1. Simulated conditions were only replicated 5 times each

Each simulated condition was replicated 5 times in WNWN and the paper reports the mean and standard deviation of the replicate clustering accuracies. The relatively small number of replicates accounts for the jerkiness of the lines in the original Figures 4 and 5 (e.g. the Bray-Curtis distances calculated on the DESeqVS normalized data). A better approach would have been to use 100 replicates as this would reduce the dependency of the results on the random number generator's seed. By increasing the number of replicates it was also possible to compare the probability of falsely and correctly clustering samples from the same and different treatment groups together (see points 9-11, below). Because the accuracies were not symmetric around the mean accuracy values the median and 95% confidence intervals or intraquartile range should have been reported. To test the effect of increasing the number of replicates, I pulled apart the code in `simulation-cluster-accuracy/simulation-cluster-accuracy-server.Rmd` into individual R and bash scripts that were executed using a Snakemake workflow with the same Conda environment I used above. This was necessary since the number of simulated conditions increased 20-fold with the additional

replicates. Such intense data processing was not practical within a single R-flavored markdown document. Again, the observed results were qualitiatively similar to those generated using the single R-flavored markdown file (Figures **pam_subsample15_fig_4.pdf** and **pam_subsample_fig_5.pdf**). The increased number of replications resulted in smoother lines and allowed me to present empirical 95% confidence intervals. For all analyses in the remainder of this paper, I used 100 randomized replicates per condition.

2. The average sizes of the libraries were small by modern standards

In the 10 years since WNWN was published, sequencing technology has advanced and sequence collections have grown considerably. For more modern datasets, it would be reasonable to expect a median number of sequences larger than 10,000 (see Table 1 of [Singleton Paper XXXXXXXX]). Therefore, I included an additional median depth of sampling value of 50,000 sequences with the original four median sequencing sampling depths (i.e., 1,000, 2,000, 5,000, 10,000). Additional sequencing coverage would be expected to result in more robust distance values since there would be more information represented in the data. Indeed, the added sampling depth showed higher accuracy values at lower effect sizes for the combinations of normalization methods and distance calculations (Figure S4 (**pam_subsample15_fig_4.pdf**)). Increased sequencing coverage also resulted in improved clustering accuracy for lower effect sizes when the library size minimum quantile was decreased (Figure S5 (**pam_subsample_fig_5.pdf**)). I will revisit the choice of the library size minimum quantile below.

3. DESeq-based variance stabilization was used with distance calculation methods that are sensitive to negative values

Close comparison of the original Figure 4 and my version (**pam_subsample15_fig_4.pdf**) revealed one important difference between the two plots. In the original analysis, the accuracies for the Weighted UniFrac distances at the largest effect size (i.e., 3.5) were 1.0 for median sequencing depths of 1,000, 2,000, and 10,000. In my version of the analysis, the values for the same sequencing depths were 0.88, 0.89, and 1.00, respectively. The 95% confidence interval for these accuracies spaned between 0.51 and 1.00. The Bray-Curtis distances were also different by both methods at smaller effect sizes and had wide confidence intervals. Inspection of the DESeq normalized OTU counts revealed that the method resulted in negative values. In fact, rendering the R-flavored markdown files in WNWN's Protocol S1 generated warning messages when passing the DESeq normalized counts to the Bray-Curtis calculator, which said, "results may be meaningless because data have negative entries in method 'bray' ". Although the Weighted UniFrac

calculator function did not generate a similar warning message, negative values would also result in similarly meaningless distances. Both are due to the fact that the calculators sum the counts of each OTU in both samples being compared. In contrast, a Euclidean distance does not use a similar sum, but sums the square of the difference between the OTU abundance in each sample. To assess the prevalence of negative counts in the simulated data, I quantified the fraction of negative values in the OTU matrix from each simulation and counted the number of simulations where the normalized OTU table had at least one negative value (Figure **deseq_negative_value.pdf**). In general the fraction of negative OTU counts increased with effect size, but decreased with sequencing effort. The fraction of simulations with at least onee negative value increased with effect size and sequencing effort. The high frequency of negative OTU counts resulted in highly variable Bray-Curtis and Weighted UniFrac values. It is likely that because the WNWN analysis only used 5 replicates that the large variation in accuracies at high effect sizes was missed initially. For the rest of this reanalysis study, I will only report results using the DESeq-based variance stabilization normalization with the Euclidean distance.

4. A single subsampling of each dataset was evaluated rather than using rarefaction, which likely resulted in noisier data

As noted above, the original jargon that was used in WNWN was confusing to many who conflated rarefying/subsampling with rarefaction. A more robust analysis would have used rarefaction since it would have averaged across random subsamplings, which individually would be unlikely to represent the overall composition of the communities. Rather than being guilty of "omission of available valid data" as claimed in WNWN, rarefaction uses all of the available data. To compare subsampling and rarefaction, I removed the 15% of samples with the lowest number of sequences for each of the 100 simulated datasets and compared the clustering accuracies from a single subsampling to rarefaction with 100 randomizations. This analysis revealed two benefits of rarefaction. First, the median distances generated by rarefaction was always at least as large as those from a single subsample (Figure **subsample_rarefaction_compare.pdf**). The difference was most pronounced for smaller average library sizes and at smaller effect sizes. The Unweighted UniFrac distances were most impacted by the use of rarefaction over subsampling. Second, the intraquartile ranges for the distances generated by rarefaction were generally smaller than those by subsampling and showed similar trends to the difference in the median distances (Figure **subsample_rarefaction_compare.pdf**). The intraquartile ranges for Bray-Curtis, Euclidean, and Unweighted UniFrac distances were actually larger by rarefaction than by subsampling at small effect sizes and average library sizes; however at larger values the subsampling intraquartile range was larger than by rarefaction for these distance calculations. Because

rarefaction incorporates more of the data and generally performed better than subsampling, the remainder of this analysis will report results using rarefaction rather than by subsampling, except when noted.

5. Results using PAM clustering were not directly compared to those of K-means and hierarchical clustering

The clustering accuracy measurements in the body of the manuscript were determined using PAM-based clusters while Protocol S1 also includes K-means and hierarchichal clustering. Although the data were not displayed in a manner that lent itself to direct comparison in Protocol S1, close inspection of the rendered figures suggested that PAM may not have been the optimal choice in all situations. Rather, K-means clustering appeared to perform better in many simulations. Because the accuracies were the smallest at lower effect sizes, I focused my comparison at the effect size of 1.15. For each set of 100 replicated simulated datasets, I compared the clustering accuracy across clustering methods to see how often each clustering method resulted in the highest accuracy (Figure **compare_cluster_methods.pdf**). Indeed, K-means clustering performed better than the other methods. Among all combinations of normalization methods, distance calculations, and read depths, PAM clustering resulted in clustering accuracies as good or better than the other methods in 49.92% of the randomizations (Figure **compare_cluster_methods.pdf**). K-means clustering was at least as good as the other methods in 74.39% of the randomizations. HClust was at least as good as the other methods for 44.32% of the randomizations. I specifically compared the clustering accuracies using rarefaction for each of the distance calculations methods using PAM and K-means clustering. Among the 30 combinations of distance calculations and read depths, K-means performed better than PAM in 29 cases with PAM doing better in the 1 other case (i.e., calculating distances with Euclidean using 10,000 sequences). When using subsampled data, K-means clustering performed better than PAM in each case. Because K-means clustering did so much better than PAM clustering in the simulated conditions, I will use K-means clustering for the remainder of this study.

6. Subsampling removed the smallest 15% of the samples, which penalized accuracy values by 15 percentage points

In WNWN, the authors quantified the tradeoff between median sequencing depth, the number of samples removed below the threshold, and clustering accuracy (original Figure 5, my Figure S5). Although the optimal threshold varied by distance metric, normalization method, and sequencing depth, they removed samples whose number of sequences was less than the 15th percentile (L404-419). They acknowledged

that this screening step, which was only used with subsampling, would decrease clustering accuracy putting it at a relative disadvantage to the other methods (page 5, column 1, last paragraph). Therefore, it was not surprising that the peak clustering accuracy for their subsampled data was at 85%. Because the true best result would not known *a priori* in an actual microbiome study, it would be impossible for researchers to conduct a sensitivity analysis comparing the tradeoffs between sequencing depth, sample number, and clustering accuracy to select a sampling depth for their analysis. The differences in clustering accuracy between subsampling and rarefaction and using PAM and K-means clustering indicated that it was necessary to reassess the tradeoff between the library size minimum quantile and clustering accuracy. When using rarefaction, K-means clustering, and only considering conditions with 2,000 or more sequences, there was not a condition where setting a higher threshold resulted in a better accuracy than using all of the samples (**kmeans_rarefaction_fig_5.pdf**). These results showed that for modern sequencing depths, using the full datasets with rarefaction and K-means clustering resulted in accuracies that were typically better than those observed when removing the smallest 15% of the samples from each simulated dataset. When the original Figure 4 was recast with these approaches, rarefaction performed at least as well as any of the other tranformations with each distance calculation, except when used with the Poisson distance (Figure **kmeans_rarefaction15_fig_4.pdf**). It is worth noting that at the largest effect sizes, K-means clustering did not perform as well as PAM for some combinations of normalization method and distance calculation (compare **kmeans_rarefaction15_fig_4.pdf** and **pam_subsample00_fig_4.pdf**); however, those combinations that performed worse by K-means were not as good as rarefaction or subsampling by either clustering method.

7. The distribution of library sizes was not typical of those commonly seen in microbiome analyses

As described above, the sequencing depths used in the 26 GlobalPatterns datasets were used as the distribution to create sequencing depths for the 80 samples that were generated in each simulation. The GlobalPatterns datasets had a mean of 1085256.8 sequences and a median of 1,106,849 sequences per dataset. The datasets ranged in sequencing depth between 58,688 and 2,357,181 sequences for a 40.16-fold difference. Rather than representing a typically observed distribution of sequencing depths that would be skewed right, the sampling distribution was normally distributed (Shapiro-Wilk test of normality, P=0.57) (Figure ***distribution_shape.pdf***). From these simulations it is unclear how sensitive the various normalizations and distance calculations were to a skewed distribution. A second limitation of this sampling distribution is that it only contained 26 unique sampling depths such that each sampling depth would have been re-used an average of 3.08 times in each simulation. Yet, it is unlikely for a real sequence collection

to have duplicate sequencing depths. To reassess the WNWN results in the context of a more typical distribution of sample sizes, I created a new set of simulations to test the effect of the shape of the distribution on the results. I created a simple sequencing depth distribution where there were 80 depths logarithmically distributed between the minimum and maximum sequencing depths of the GlobalPatterns dataset (Figure ***distribution_shape.pdf***). The median of this distribution was 372,040 and the mean was 629824.8. When I regenerated the original Figures 4 and 5 using the log-distributed sequencing effort distribution, the differences in normalization methods were more apparent (Figure ***fig_4_kmeans_rarefaction00_log.pdf***). For each of the distance calculators, rarefaction to the size of the smallest dataset yielded accuracies that were at least as good as the other methods across effect sizes and median sequencing depths. The difference was most pronounced at smaller effect sizes and sequencing depths. When comparing the performance of rarefaction across distance calculators for different effect sizes, sequencing depths and size of smallest sample (***fig_5_kmeans_rarefaction_log.pdf***), the accuracies I observed using the log-distributed sample sizes was at least as good as those obtained using the GlobalPatterns-based distribution ***fig_4_kmeans_rarefaction00_a.pdf***. The issue of the number of samples in a study and the distribution of their sequencing depths in the context of controlling for uneven sampling effort is explored in far greater detail in an another analysis using sequencing depths observed in actual biological samples [XXXXXX].

8. A filtering step was applied to remove rare taxa from the simulated datasets

McMurdie and Holmes were emphatic that "**rarefying biological count data is statistically inadmissible** because it requires the omission of available valid data" (emphasis in original). Thus it is strange that they argue against removing data when rarefying/subsampling, but accept removing rare and low-prevalence OTUs prior to normalizing their counts. This practice has become common in microbiome studies and is the standard approach in tools such as dada2, unoise, and deblur [XXXXXXXX]. However, my previous work has shown that rare sequences from a poorly sequenced sample often appear in more deeply sequened samples suggesting that they are not necessarily artifacts. Furthermore, removing rare sequences alters the structure of communities and has undesirable effects on downstream analyses [XXXXXXXX]. Although my previous work does an extensive analysis of the effects of removing rare sequences, I wanted to explore the effect of filtering in the context of the WNWN simulation framework. For each of the filtered and non-filtered OTU tables I calcualted the absolute value of the difference in accuracy between each distance calculation following the normalization procedure (Figure **compare_filter_accuracy.pdf**). With the exception of the Weighted UniFrac distances, each of the distance calculations and normalization procedures were sensitive to the filtering. The rarefaction data tended to be sensitive to filtering at small effect sizes.

Distances generated using raw counts, DESeq variance stabilization, and Upper Quartile Log Fold Change tended to be more sensitive to filtering at larger effect sizes. When using relative abundance data, Bray-Curtis distances were sensitive to filtering at small effect sizes and Unweighted UniFrac distances were sensitive at large effect sizes. These trends appeared to be driven by the dependence of the distance calculation on low abundance taxa. More surprising than the effect of filtering on the mean absolute difference in clustering accuracy was the wide variation in accuracies at each effect size. Among the different normalization methods, the accuracies calculated using rarefaction had the narrowest 95% confidence interval for all distance calculations except for calculating Unweighted UniFrac distances. For these distances, using relative abundances had the narrowest range at small effect sizes; rarefaction had the narrowest range at larger effect sizes. Again, these trends appeared to be driven by the dependence on low frequency taxa in Unweighted UniFrac, which is dependent on the presence or absence of taxa rather than their abundance. Given my previous work and the large variation caused by removing rare taxa, OTU filtering should not be performed in microbiome analyses.

9. No accounting for difference in performance when library sizes are confounded with treatment group

In previous analyses I have observed that not using rarefaction can lead to falsely detecing differences between communities when sampling effort is confounded with the treatment group [XXXXXXX]. Such situations have been observed when comparing communities at different body parts where one site is more likely to generate contaminating sequence reads from the host [XXXXXXX]. My previous analyses showed that rarefaction did the best job of controlling the rates of false detection (i.e., Type I errors) and maintaining the statistical power to detect differences (i.e., 1-rate of Type II errors) of differences between groups of samples. To determine whether this result was replicated with the WNWN simulation framework, I created a skewed sampling distribution using both the GlobalPatterns and Log-distributed sequence distributions. To skew the sample counts the sequencing depth of samples from one treatment group were drawn from below the median number of sequences of the sampling distribution and those for the second treatment group were from above the median. To assess Type I errors, I compared the clustering accuracies using an effect size of 1.0 using both the skewed and unskewed sampling distributions (Figure **cluster_skew_compare_i.pdf**). The samples should have only been assigned to one cluster; however, each of the clustering methods forced the samples into two clusters. So, when there are two groups of 40 samples that do not differ, the best a method could do would be to correctly assign 41 of the 80 samples for an accuracy of 0.51. The Type I error did not vary by method when the sequencing depth was not skewed. Yet, when the sequencing

depth was skewed, rarefaction was the most consistent normalization method for controlling Type I errors. At larger effect sizes the power to detect differences increased when the sequencing depth was skewed (Figure **cluster_skew_compare_ii.pdf**). At the effect size of 1.15, the rarefied data generated the highest accuracy clusters regardless of whether the data were skewed. The exception to this were the Poisson distance clusters, which generally clustered performed poorly. Although the skewed simulation is extreme, it highlights the ability of rarefaction to control Type I errors while maintaining high power and the sensitivity of the other normalization methods.

10. Clustering accuracy was used rather than direct comparisons of beta diversity

Since WNWN was published, there has been controversy over the use of clustering methods to group samples (i.e., enterotypes). Concerns have been raised including whether such clustering should be done on ecological distances or sequence counts and the biological interpretation of such clusters [XXXXXXXXX]. As described in the previous point, one notable challenge with using clustering accuracy as the dependent variable is that the clustering methods force the samples into one of two clusters. For the case where the effect size was 1.0, it was impossible for all 80 samples to be assigned to a single cluster. As has already been described in point 5, an additional problem with clustering is the variation in the relative performance of a method across conditions. A more commonly used approach for analyzing distance matrices is to use a non-parametric analysis of variance test of the various distance matrices (i.e., AMOVA, PERMANOVA, NP-ANOVA)[XXXXXXXXXXX]. I subjected each of the distance matrices to such a test using `adonis2`, a function from the vegan R package that implements this test to assess the effects of each normalization and distance calculation method on the Type I errors and statistical power. As was seen above, when sequencing depths were randomly distributed across the two treatment groups, the Type I error did not meaningfully deviate from the expected 5% (Figure **adonis_skew_compare_i.pdf**). Again, when sequencing depths were skewed between the two treatment groups, rarefaction was the only normalization approach to control the Type I error. Similar to the clustering accuracy results, when distances were calculated using rarefaction, the tests consistently had the best statistical power (Figure **adonis_skew_compare_ii.pdf**). When considering both Type I error and power, rarefaction performed the best among the different normalizations.

11. There was no consideration of effects of normalization methods on alpha diversity metrics

Rarefaction was originally proposed as a method for controlling uneven sampling effort when comparing community richness values [XXXXXXXXXXXXX]. Thus it was surprising that WNWN did not consider

the effect of the proposed normalizations on alpha-diversity metrics such as richness or Shannon diversity. Therefore, for each of the normalizations, I compared the richness and diversity of the two treatment groups. The DESeq normalized data were not included because the normalization produced negative values which were not compatible with calculations of richness or Shannon diversity. Also, data from the Upper Quartile Log Fold Change normalization were not used for richness calculations since the normalization returned the same richness values for each sample regardless of the treatment group. I assessed significance for each iteration using the non-parametric Wilcoxon two-sampled test. I compared the risk of committing Type I errors and the power to detect differences by the different normalizations (Figure **alpha_compare.pdf**). For these analyses, I used the GlobalPatterns data with the random and skewed distribution of samples. Similar to the results in points 9 and 10, with the exception of rarefaction, the simulations using a skewed distribution resulted in all of the replicates having a significant test with each of the other tranformations. The power to detect differences in richness and diversity at effect sizes of 1.15 and greater with rarefaction was at least as high as any of the other normalizations.

One odd result from this analysis was that the power to detect differences in richness at small effect sizes increased between 1,000 to 2,000 sequences with values of 91.00 and 100.00, respectively. The power then decreased with increasing sequencing effort to 57.00 with 50,000 sequences. This appeared to be because although the parent distributions had very different shapes, they shared a large number of rare taxa. When using rarefaction to compare the distributions at the size of the smallest distribution (i.e., 3,598,077 sequences), the Feces parent distribution had 1,559.65 OTUs and the Ocean had 1,335.00 OTUs; they shared 894.66 OTUs. However, when comparing the distributions at 1,000 and 50,000 sequences, the Ocean distribution had greater richness than the Feces distribution by 43.76 and 6.05, respectively. Therefore, although more replicates at lower effect sizes yielded a small p-value at shallow rather than deeper sequencing depths for richness, the direction of the difference in richness was incorrect. This result underscores the challenges of using presence-absence based-metrics like richness and Jaccard and Unweighted UniFrac distances to compare microbial communities.

Discussion

The conclusions from McMurdie and Holmes's study have had a lasting impact on how researchers analyze microbiome sequence data. As I have demonstrated using their original simulation approach their claims are not supported. The most important points that lead to the difference in our conclusions include the choice of clustering algorithm and arbitrarily selecting a sequencing effort threshold that was used to re-

move samples. Furthermore the decision to evaluate the effectiveness of normalization methods based on clustering samples adds a layer of analysis that has been controversial and not widely used. Ultimately, the authors choice of the word “rarefying” has sewn confusion in the field because it is often used in place of “rarefaction”. As I have demonstrated there were numerous choices throughout the orininal study that made rarefying/rarefaction look worse than it truly was. In fact, when the data are taken as a whole, rarefaction is the preferred approach. Short of obtaining the exact same number of sequences from each sample, rarefaction remains the best approach to control for uneven sampling effort when analyzing alpha and beta diversity metrics.

Beyond the discussion of whether rarefaction is appropriate for analyzing microbiome data it is worth commenting on McMurdie and Holmes’s advice to use DESeq’s Variance Stabilization or edgeR’s Upper Quartile Log-fold Change normalization strategies. These methods have been adopted from gene expression analysis to microbiome analysis. Gene expression analysis implicitly assumes that all samples have the same genes since. While this might work in comparing healthy and diseased tissues from a cohort of patients, it does not generalize to those patients’ microbiota. Microbial populations are highly patchy in their distribution. Thus, a zero count for gene expression is more likely to represent a gene below the limit of detection whereas a zero count for a microbiome analysis is more likely to represent the true absense of the OTU. An example of where this is relevant is the necessity of adding a pseudocount to all OTUs to perform both the edgeR and DESeq-based normalizations (L443 and L487, respectively). In WNWN, a pseudocount of 1 was used. However, this value is arbitrary and the sensitivity of the results can vary based on the patchiness of the communties being analyzed. Since WNWN was published, compositional approaches have been proposed to account for uneven sampling and to provide improved interpretability [XXXXXX]. However, these methods also often require the use of pseudocounts and are not actually insensitive to uneven sampling [XXXXXXXX]. Rarefaction is preferred to these alternatives.

The choice of distance metric is a complicated question and the use of six different metrics in WNWN illustrates the challenges. Within the ecology literature, Euclidean distances are widely avoided because joint absense is weighted the same as joint presence of taxa [XXXXXX]. As discussed in point 11, metrics that are based on community membership (i.e., the presence or absense of taxa) performed worse than those than those that were based on community structure (i.e., their relative abundance). For this reason, the Unweighted UniFrac and other metrics like the Jaccard or Sorenson distance coefficients should likely be avoided. The Poisson distance metric is largely novel to the microbial ecology literature and performed no better than the more traditional metrics. In the current analysis, the phylogenetic Weighted UniFrac distance performed comparably to Bray-Cutis distances for clustering or differentiating between communities

with adonis. In practice, however, the current reality is that it is computationally impractical to construct phylogenetic trees with modern datasets. Although algorithms have been developed to sidestep *de novo* construction of trees [XXXXXX], these algorithms depend on reference-based clustering strategies that have significant challenges [XXXXXX]. Among the options analyzed here, Bray-Curtis is the most robust and practical choice. Regardless, all of the methods perform best when using rarefaction relative to the other normalization methods.

Although the authors claim that "Rarefying counts requires an arbitrary selection of a library size minimum that affects downstream inference" (page 8, column 1, point 3), in acutal microbiome studies the selection of a sampling depth is not as arbitrary as the authors claim. Rather, to avoid "p-hacking", researchers pick a set of criteria where they will include or exclude samples prior to testing their data. Examples of criteria might include the presence of a large gap in the sequencing effort distribution, a desire to include poorly sequenced controls, or the *a priori* stipulation of a minimum sequencing effort. To mitigate concerns of arbitrary or engineered minimum library sizes, researchers should indicate the rationale for the threshold they selected.

WNWN performed a second set of simulations to address the effect of normalization method on the ability to correctly detect differential abundance of OTUs that were randomly selected to have their relative abundances changed (i.e., Simulation B). Re-addresssing this set of simulations is beyond the scope of the current anaylsis and others have already contributed critiques [XXXXXXX]. However, many of the same concerns addressed above would apply. Perhaps more importantly is the fact that if the relative abundance of several OTUs increase, then the relative abundance of all other OTUs would necessarily decrease. Thus, one would expect every OTU to be differentially abundant. Because of this, it is not possible to truly modify the abundance of a set of OTUs indepdent of all othter OTUs. This is an important limitation of tests of methods attempting to detect differentially abundant OTUs. Regardless, a form of rarefaction could still be employed for detecting differential abundance. One could subsmaple the data, perform the statistical test, and identify the differentially abundant OTUs. This process could be repeated. In my experience the largest differences between subsamples are for low relative abundances OTUs, which are unlikely to be biologically or statistically significant.

In a parallel set of analyses I have used alternative simulation and evaluation strategies to look closer at rarefaction and its alternatives and the practice of filtering low abundance seqeunces. The results of those studies are similar to this reanalysis. These studies demonstrate that sequence and OTU filtering "requires the omission of available valid data" and that rarefaction is the only available method for controlling the effects of uneven sampling. Far from being inadmissible, rarefaction of unfiltered datasets yields the most

robust results.

Methods

Code availability

A git repository containing all of the code needed to reproduce this study is available at https://www.github.com/SchlossLab/Schloss_WNWN_MSpectrum_2023. All simulations and analyses were performed using R and bash scripts using Snakemake (v7.24.0) to track dependencies and automate the pipeline and conda (v4.12.0) and mamba (v1.1.0) to specify software and package versions (see workflow/envs/nr-base.yml in repository). This manuscript was written as an R-flavored markdown document and rendered using the `rmarkdown` R package (v2.18). All figures were generated using `dplyr` (v1.1.0) and `ggplot2` (v3.4.2) from the `tidyverse` metapackage (v1.3.2) and `ggtext` (v0.1.2) and `ggh4x` (v0.2.4) within R (v.4.2.2; see workflow/envs/nr-modern.yml in repository).

Reproducing WNWN Protocol S1

The compressed directory published with WNWN as Protocol S1 is available on the PLOS Computational Biology website with the original paper at https://doi.org/10.1371/journal.pcbi.1003531.s001. To render the R-flavored markdown files to HTML files using software and packages as similar as possible to that of the original study, I created an environment (see workflow/envs/nr-s1.yml in repository) that contained versions as close to those indicated in the pre-rendered files. The most important packages for the reproduction of the cluster analysis included `DESeq` (my version: v1.39.0 vs. WNWN version: v1.14.0), `edgeR` (v3.30.0 vs. v3.4.0), `cluster` (v2.1.0 vs v1.14.4), `phyloseq` (v1.32.0 vs. v1.6.0), and `PoiClaClu` (v1.0.2.1 vs. 1.0.2). In addition, I used R v4.0.2 whereas the original WNWN analysis used v3.0.2. Finally, to render the `simulation-cluster-accuracy-server.Rmd` file to HTML, it was necessary to apply a patch to the code. This patch set the path to the location where the figures should be saved and removed the R code that deleted all objects in the enviornment. Both changes were necessary to better organize the project and had no bearing on the content of the HTML file that was rendered.

Simulated communities

The code contained within `simulation-cluster-accuracy-server.Rmd` was spilt into individual R scripts and modified to make the generation of the simulated communities more modular and scalable. The code

for geneating the simulated communities was run using the nr-s1 conda environment. The Ocean and Feces parent distributions were generated as described above. Five variables were altered to simulate each set of communities. First, eight mixing fractions were used to manipulate the effect size between the two simulated treatment groups with 40 samples per group. These were the same values used in the original WNWN study (1.0 [i.e., no difference], 1.15, 1.25, 1.5, 1.75, 2, 2.5, and 3.5). Second, the number of sequences in each of the 80 samples was either randomly selected from the 26 library sizes contained wihtin the GlobalPatterns dataset or from a log distribution of 80 sequencing depths evenly distributed between the smallest (58,688 sequences) and largest dataset (2,357,181 sequences) in the GlobalPatterns dataset (Figure **distribution_shape.pdf**). Third, the resulting 80 sequencing depths were scaled so that the median sequencing depth across the 80 samples was either 1,000, 2,000, 5,000, 10,000, and 50,000 sequences. Fourth, the sequencing depth of each sample was either randomly assigned to each treatment group or assigned so that those depths less than or greater than the median were assigned to separate treatment groups. Finally, each set of conditions was replicated 100 times. These five parameters resulted in 1,600 simulated datasets (8 fractions x 2 sequencing depth models x 5 sequencing depths x 2 sample assignment models x 100 replicates).

Generation of distances between communities

Next, each simulation was further processed by filtering rare OTUs, to normalize uneven sampling depths, and to calculate ecological distances between samples. Again, the code contained within `simulation-cluster-accuracy-server.Rmd` was spilt into individual R scripts and modified to make the generation of the simulated communities more modular and scalable within the nr-s1 conda environment. First, each simulation was either filtered to remove rare OTUs or left unfilterd. As was done in the original WNWN study, filtering was performed in two steps: (1) any OTUs whose total abundance was less than 3 across all 80 samples and that did not appear in at least 3 samples and (2) any OTU that did not have more than 1 sequence in more than 5% of the 80 samples (i.e., 4 samples) and that did not have a total abundance across the 80 samples greater than one half of the number of samples in each community type (i.e., 20) were removed. Second, OTU data was normalized by one of six approaches: (1) non-normalized raw counts, (2) relative abundance, (3) varaiance stabilization using DESeq, (4) Upper Quartile normalization using edgeR, (5) a single subsampling to a common number of sequences, (6) rarefaction with 100 randomizatoins to a common number of sequences. Although the data were not presented, the original and my code also normalized the counts using trimmed mean of M-values (TMM) and relative log expression (RLE) in edgeR. The level of subsampling and rarefaction was selected based

on the size of the sample whose sequencing depth was at the 0 (i.e., included all of the samples), 5, 10, 15, 20, 25, and 40th percentiles. Finally, the normalized simulated datasets were used to calculate pairwise distances between the 80 samples using seven different calculations: Bray-Curtis, Euclidean, Poisson, Mean Squared Difference, Unweighted UniFrac, Weighted Unifrac, and Biological Coeffcient of Variation (BCV); however, data using BCV were not discussed in this or the original analysis. For the rarefaction normalized data, each of the 100 subsampled datasets were used to calculate a distance matrix. The mean of the 100 distance matrices was used as the distance matrix for the rarefaction data. As indicated in the Results section, there were combinations of normalization and distance methods that were not compatible (e.g., DESeq Variance Stabilization normalization with Bray-Curtis distances). In such cases, distance matrices were coded as `NA`. My choice of combinations of normalziations and distances to include was determined by the original analysis except where noted. For each simulated dataset there were 280 possible combinations of filtering, normalization, and distance methods (2 filtering methods x 20 normalization method x 7 distance methods).

Analysis of distances between communities

Two general strategies were used to analyze the pairwise distances in each distance matrix. First, by recycling the the code contained within `simulation-cluster-accuracy-server.Rmd` in the nr-s1 conda environment, samples were assigned to one of two clusters using partitioning around the medoid (PAM) with the `pam` function from the `cluster` package, K-means clustering with the `kmeans` function from the `stats` base package, and hierarchical clustering with the `hclust` and `cutree` functions from the `stats` base package. The accuracy of the clustering was measured as the fraction of the 80 samples assigned to the correct cluster. As in the original analysis, if samples were removed because the number of sequences in them fell below the threshold for subsamplnig or rarefaction, then the accuracy could be below 50%. For the 1 mixing fraction all of the samples should have been assigned to one group; however, because they were forced into two groups there would be 41 "correctly" assigned samples of the 80 (51.25%). Second, the distance matrices were analyzed for significant different centroids using the `adonis2` function from the `vegan` package (v2.6-4) in the nr-modern environment. The fraction of significant tests was the fraction of the 100 replicate simulations that had a p-value less than or equal to 0.05. The false positive rate for a condition was the fraction of significant tests when the mixing level was 1. The power was the fraction of significant tests at other mixing levels.

Analysis of richness and diversity between communities

The original WNWN analysis did not address the impact of normalization on alpha diversitiy metrics. Using the nr-s1 conda environment, I measured the richness and Shannon diversity using the normalized OTU counts. Richness was measured by counting the number of OTUs for each sample across the simulations and the Shannon diversity using their relative abundance using the commonly used forumla. For the rarefaction normalized data, each of the 100 subsampled datasets were used to calculate the richness and diversity. The mean of the 100 values was used as the value for the rarefaction data. To assess differences between the two treatment groups, the richness and diversity values were compared using the two-sample Wilcoxon non-parametric test with the `wilcox.test` from the `stats` base package. Similar to the adonis analysis, the fraction of significant tests was the fraction of the 100 replicate simulations that had a p-value less than or equal to 0.05. The false positive rate for a condition was the fraction of significant tests when the mixing level was 1. The power was the fraction of significant tests at other mixing levels.

Acknowledgements

I am grateful to Paul McMurdie and Susan Holmes for publishing the source code that they used to conduct their analysis as R-flavored markdown files in the supplement to WNWN. This provided me with a better understanding of their methods including noticing that only a single subsampling step was performed for each random seed and that only 5 random seeds were used. Furthermore, their code enabled me to replicate and build upon their simulations.

References

Figures