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

Running title: Revisiting “Waste not, want not”

Patrick D. Schloss†

† To whom correspondence should be addressed:

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Research article

Abstract

In 2014, McMurdie and Holmes published the provocatively titled, “Waste not, want not: why rarefying microbiome data is inadmissible”. The claims of their study have significantly altered how microbiome researchers control for the unavoidable uneven sequencing depths that are inherent in modern 16S rRNA gene sequencing. Confusion over the distinction between the definitions of rarefying and rarefaction continues to cloud the interpretation of their results. More importantly, the authors made a variety of problematic choices when designing and analyzing their simulations. I identified 11 factors that could have compromised the results of the original study. I reproduced the original simulation results and assessed the impact of those factors on the underlying conclusion that rarefying data is inadmissible. Throughout, the design of the original study made choices that caused rarefying and rarefaction to appear to perform worse than they truly did. Most important were the approaches used to assess ecological distances, the removal of samples with low sequencing depth, and not accounting for conditions where sequencing effort is confounded with treatment group. Although the original study criticized rarefying for the arbitrary removal of valid data, repeatedly rarefying data many times (i.e., rarefaction) incorporates all the data. In contrast, it is the removal of rare taxa that would appear to remove valid data. Overall, I show that rarefaction is the most robust approach to control for uneven sequencing effort when considered across a variety of alpha and beta diversity metrics.

Importance

Over the past 10 years, the best method for normalizing the sequencing depth of samples characterized by 16S rRNA gene sequencing has been contentious. An often cited paper by McMurdie and Holmes forcefully argued that rarefying the number of sequence counts was “inadmissible” and should not be employed. However, I identified a number of problems with the design of their simulations and analysis that compromised their results. In fact, when I reproduced and expanded upon their analysis, it was clear that rarefaction was actually the most robust approach for controlling for uneven sequencing effort across samples. Rarefaction limits the rate of falsely detecting and rejecting differences between treatment groups. Far from being “inadmissible”, rarefaction is a valuable tool for analyzing microbiome sequence data.

Introduction

Microbiome studies that use amplicon sequencing to characterize multiple samples use PCR to amplify 16S rRNA gene fragments using primers with distinct barcodes or index sequences (1–3). These barcodes allow researchers to pool PCR products and then deconvolute the resulting sequence data based on the barcode sequences. Despite researchers' best efforts to generate equimolar pools of PCR products, it is common to observe more than 10-fold variation in the number of sequences per sample (4). Researchers desire strategies to minimize uneven sequencing depth and thus need methods to control for this unevenness in their analyses. Of course, uneven sampling effort is not unique to microbiome research and is a challenge faced by all community ecologists (5, 6). Common approaches to control for the effects of uneven sequencing depths include use of proportional abundance (i.e., relative abundance), scaling of counts, parameter estimation, and rarefaction (7–23).

In 2014 Paul McMurdie and Susan Holmes published "Waste not, want not: why rarefying microbiome data is inadmissible" (WNWN) in *PLOS Computational Biology* (24). The provocative title attempts to express the idea that if one does not waste resources then they will not be found wanting the resource when it is needed. Applied to their study, it implies that rarefying data arbitrarily discards samples and sequence data in the subsampling step leading to a loss of statistical power. Their 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 2,560 times as of October 2023. There has been a rebuttal of WNWN that showed how rarefaction is beneficial in some cases (25); however, the proponents of WNWN appear to be holding sway over the microbiome community. 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. I have also received similar comments from reviewers regarding my own work. In responding to such critiques, I have grown to appreciate that there is significant confusion in the field over what is meant by "rarefying" and "rarefaction".

It was unfortunate that McMurdie and Holmes used the term "rarefying" throughout their manuscript. The authors were correct to state that the distinction between "rarefying" and "rarefaction" is confusing and leads to their conflation. They defined rarefying as taking a subsample of the same number of reads from each sample ($N_{L,m}$) without replacement and discarding any samples that have fewer than that number of reads. The subsampling is performed without replacement. Traditionally, rarefaction involves repeating the subsampling step a large number of times, calculating a metric, and averaging the metric across the subsamples (12, 13, 26). In other words, rarefying or subsampling is rarefaction, but with a single randomization.

Confusion over the terms is demonstrated in the choice of citations that McMurdie and Holmes used to define rarefying (i.e., (13, 27, 28)). The cited studies either did not use the words “rarefy” or “rarefying” or used them interchangeably with rarefy as a verb form of rarefaction. For example, Hughes and Hellmann did not use “rarefy” (13). Rather they used “rarefaction” in the traditional sense with multiple subsamplings. Meanwhile, the QIIME-based literature appears to use “rarefy” and “rarefaction” interchangeably to mean only a single subsampling (27, 28). Confusion comes from the WNWN authors’ admonition that “[i]n many cases researchers have also failed to repeat the random subsampling step”. This seems to call on researchers to use rarefaction rather than the single subsampling step. Subsequent researchers have continued to conflate the terms when citing WNWN (**Supplemental Text**). An exemplar of the confusion is the creation of a technique that uses “repeatedly rarefying” as an approach distinct from rarefaction when they were in fact re-proposing traditional rarefaction (29).

To minimize confusion, I will use “subsampling” in place of “rarefying” through the remainder of this study and I will use 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 (typically 100 or 1,000). Researchers are encouraged to report the number of iterations.
6. Compute summary statistics (e.g., the mean) using the values from step 4.

This definition aligns with how rarefaction was originally defined for comparing richness (i.e., the number of taxa in a community) across communities when communities are sequenced to different depths (5, 6). With this more general approach to the definition of 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`, `phylo.diversity`, and `dist.shared`. The procedure outlined above could also be used for hypothesis tests of differential abundance; however, further consideration is needed to synthesize the results of these tests across a large number of replications.

Description and critique of “Simulation A” from WNWN

McMurdie and Holmes analyzed the effect of subsampling and other approaches on clustering accuracy using what they called “Simulation A” in their Figure 2A and elsewhere in WNWN. 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 (R markdown) file that was published as Protocol S1 in WNWN. I will outline the simulation strategy below and will reference line numbers from their R markdown file with “L” as a prefix.

For the WNWN cluster analysis, OTU abundances and sequencing depths were obtained from the GobalPatterns dataset (2), which consisted of 26 samples originally obtained from 9 environments including creek (n=3), human feces (n=4), human tongue (n=2), lake (n=2), mock communities (n=3), ocean (n=3), sediment (n=3), skin (n=3), and soil (n=3). In the WNWN analysis, the 7 human fecal and ocean sequence datasets were used to generate the OTU distributions for the two simulated treatment groups (L129). To generate the fecal and ocean distributions, 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 obtained the identifiers of the first 2,000 OTUs (L66). Returning to the 7 samples they selected the data for the corresponding 2,000 OTUs and pooled the OTU abundances of the fecal and ocean samples separately to create two OTU abundance distributions (L144, L159-160, L197-198). Next, the fecal and ocean distributions were mixed in 8 different fractions to generate two community types that differed by varying effect sizes (i.e., 1.00, 1.15, 1.25, 1.50, 1.75, 2.00, 2.50, and 3.50; L170-195, L220); an effect size of 1.00 generated a null model with no difference between the treatment groups. To simulate the variation in sequencing depth across the 80 samples, they normalized the number of sequences from each of the 26 samples in the full GlobalPatterns dataset so that the median number of sequences ($\tilde{N}_L$) across the samples had 1,000, 2,000, 5,000, or 10,000 sequences (L324-325). They then randomly sampled the 26 normalized sequencing depths, with replacement, to generate 80 sequencing 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). 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). Simulation A consisted of 160 simulations (8 effect sizes x 4 median sequencing depths x 5 replicates = 160 simulations).

After generating the OTU counts for the 160 simulations, the authors applied several normalization methods, distance calculations, and clustering algorithms to the data. To normalize the OTU data the WNWN analysis either applied no normalization procedure (L559), calculated OTU relative abundances (L392-401, L562-563), subsampled the data (L409-416, L566-569), performed Variance Stabilization using the `DESeq` R package (L478-504, L580-583)(7), or performed Upper-Quartile Log-Fold Change normalization using the `edgeR` R package (L425-457, L576-577, L622-721) (19). 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 (L96, L1065-1077).

Non-normalized data were used to calculate Bray-Curtis, Euclidean, Poisson, and Weighted UniFrac distances (L793-798, L801-806). Relative abundance data were used to calculate Bray-Curtis, Unweighted UniFrac , and Weighted UniFrac distances (L760-767). Subsampled data were used as input to calculate distances between samples using Bray-Curtis, Euclidean, Unweighted UniFrac, Weighted UniFrac distances, Poisson distances, and top-Mean Squared Difference (L769-775, L793-798, L801-806). DESeq Variance Stabilization normalized data were used to calculate Bray-Curtis, Euclidean, and Weighted Unifrac distances (L793-798). The Upper-Quartile Log-Fold Change normalized data were only used to calculate top-Mean Squared Difference distances (L801-806). WNWN calculated Bray-Curtis (30), Euclidean (30), Unweighted UniFrac (31), and Weighted UniFrac distances (32) using the `Phyloseq` R package (33). They calculated Poisson distances using the `PoiClaClu` R package (34). The top-Mean Squared Difference was calculated using the `edgeR` R package (19). 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 (L865-879). Although data for all three methods were presented in Protocol S1, only the PAM data were 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 (L887-908). Since some of the subsampling conditions removed samples below a minimum sequencing depth threshold, the removed samples were counted as mis-clustered samples yielding minimum accuracies below 50%.

Although all simulations represent an artificial representation of reality and can be critiqued, I have identified eleven elements of the design of Simulation A that warranted further review.

1. Simulated conditions were only replicated 5 times each, potentially increasing the sensitivity of results to the choice of the random number generator seed
2. The average sizes of the libraries do not cover the larger sequencing depths frequently found in modern microbiome studies, which may limit the generalizability of the results
3. DESeq-based Variance Stabilization generates negative values and was used with distance calculation methods that are sensitive to negative values, which likely led to nonsense distances and clusters
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, although close inspection suggests that K-means may have been superior to PAM for some conditions
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, which may limit the generalizability of the results
8. A filtering step was applied to remove rare taxa from the simulated datasets, which may have skewed the shape of the community distributions
9. There was no accounting for differences in performance when library sizes are confounded with treatment group
10. Clustering accuracy was used rather than the more direct and frequently applied comparisons of beta diversity using permutation tests
11. There was no consideration of effects of normalization methods on richness, which is the traditional application of rarefaction

Below, I replicated the original WNWN simulations and evaluated these points 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 WNWN using the authors’ code. I created a Conda environment that used the R version and package versions that were as close as possible to those used in WNWN. It was necessary to patch the WNWN's R markdown file to render the document to be compatible with the overall workflow of this study. I was able to generate figures similar to those presented as WNWN's Figures 4 and 5; my results are shown in this study as Figures S1 and S2, 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 markdown file (i.e., Protocol S1 labels the figures as Figures 3 and 4 corresponding to the published Figures 4 and 5). Regardless of the differences, my results were qualitatively similar to that provided in WNWN's Protocol S1.

1. Simulated conditions were only replicated 5 times each, potentially increasing the sensitivity of results to the choice of the random number generator seed

Each simulated condition was replicated 5 times in WNWN and the paper reported the mean and standard deviation of the replicate clustering accuracies. The relatively small number of replicates accounts for the jerkiness of the lines in WNWN's Figures 4 and 5 (e.g. the Bray-Curtis distances calculated on the DESeq Variance Stabilization 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 10 and 11, below). Because the accuracies were not symmetric around the mean accuracy values, the median and 95% confidence intervals or interquartile 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 markdown document. Again, my results were qualitatively similar to those generated using the single R markdown file (Figures S3 and S4). 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 do not cover the larger sequencing depths frequently found in modern microbiome studies, which may limit the generalizability of the results

In the 10 years since WNWN was published, sequencing technology has advanced and sequence collections have grown considerably. Although sequencing depths at the smaller end of the range used in WNWN are reported in recent studies, it is increasingly common to find a median number of sequences larger than 10,000 (see Table 1 of (4)). Therefore, I included an additional median depth of sequencing of 50,000 sequences with WNWN's four median sequencing 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 sequencing depth showed higher accuracy values at lower effect sizes for the combinations of normalization methods and distance calculations (Figure S3). Increased sequencing coverage also resulted in improved clustering accuracy for lower effect sizes when the library size minimum quantile was decreased (Figure S4). I will revisit the choice of the library size minimum quantile below.

3. DESeq-based Variance Stabilization generates negative values and was used with distance calculation methods that are sensitive to negative values, which likely led to nonsense distances and clusters

Close comparison of WNWN's Figure 4 (Figure S1) and my version (Figure S3) revealed several difference between the two plots. First, in the WNWN analysis, the accuracies for the Weighted UniFrac distances calculated using DESeq Variance Stabilization normalization at the largest effect size (i.e., 3.5) were 1.00 with no variation between replicates for median sequencing depths of 1,000, 2,000, and 10,000. In my version of the analysis, the median values for the same sequencing depths were also 1.00. However, the 95% confidence intervals spanned between 0.51 and 1.00 indicating a considerable amount of variation in the accuracy values. Second, clustering accuracy for Bray-Curtis distances also calculated using DESeq Variance Stabilization normalized OTU counts were different between the original and my simulation at smaller effect sizes and had wide confidence intervals. Inspection of the DESeq Variance Stabilization normalized OTU counts revealed that the method resulted in negative values. It has been suggested that WNWW turned negative DESeq normalized counts to zero (25); however, I was unable to find code in `simulation-cluster-accuracy-server.Rmd` that made this transformation. In fact, rendering the R 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 count values would also result in similarly meaningless distances. Both are due to the fact that the distance 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. Even if negative values were converted to zeroes, this would effectively the same as removing rare taxa, which could have a significant impact on the shape of the communities (4). 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 S5). In general the fraction of negative OTU counts increased with effect size, but decreased with sequencing effort. The fraction of simulations with at least one 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

A more robust analysis would have used rarefaction since it would have averaged across a large number of random subsamplings (e.g., 100 or 1,000). By using a large number of subsamplings the likelihood of incorporating all of the OTUs would have increased. Rather than being guilty of “omission of available valid data” as claimed in WNWN, with a sufficient number of subsamplings, traditional rarefaction uses all of the available data. To fairly compare subsampling, as employed in WNWN, and rarefaction, I removed the 15% of samples with the lowest number of sequences 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 1). 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 interquartile ranges in clustering accuracy by rarefaction were generally smaller than those by subsampling and showed similar trends to the difference in the median distances (Figure 1). 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, although close inspection suggests that K-means may have been superior to PAM for some conditions

The clustering accuracy measurements reported in the body of WNWN were determined using PAM-based clusters while Protocol S1 also includes K-means and hierarchical 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. Actually, 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 2). Indeed, K-means clustering performed better than the other methods. Among all combinations of normalization methods, distance calculations, and read depths, K-means clustering was at least as good as the other methods in 74.39% of the randomizations (Figure 2); PAM clustering resulted in clustering accuracies as good or better than the other methods in 49.92% of the randomizations; and hierarchical clustering (HClust) was at least as good as the other methods for 44.32% of the randomizations. Finally, 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 used 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 (WNWN's Figure 5, my Figure S4). 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 threshold would not be 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 sequencing depth for their analysis without the risk of p-hacking. The differences in clustering accuracy between subsampling and rarefaction with 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 (Figure 3). These results showed that for modern sequencing depths, using the full datasets with rarefaction and K-means clustering resulted in accuracies that were better than those observed when removing the smallest 15% of the samples from each simulated dataset. When the WNWN Figure 4 was recast with these approaches, rarefaction performed at least as well as any of the other transformations with each distance calculation (Figure 4).

7. The distribution of library sizes was not typical of those commonly seen in microbiome analyses, which may limit the generalizability of the results

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 1,085,256.8 sequences and a median of 1,106,849 sequences per dataset (Figure S6). 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 (see Figure S1 from (4)), the sequencing depth distribution was normally distributed (Shapiro-Wilk test of normality, P=0.57). From these simulations it is unclear how sensitive the various normalizations and distance calculations were to a more realistic skewed distribution. A second limitation of this sequencing depth distribution is that it only contained 26 unique sequencing depths such that each sequencing 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 sequencing depths, 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 S6). The median of this distribution was 372,040 and the mean was 629,824.8. When I regenerated the WNWN's Figures 4 and 5 using the log-scaled sequencing effort distribution, the differences in normalization methods were more apparent (Figure 5 and S7). 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, the accuracies I observed using the log-scaled sample sizes was at least as good as those I obtained using the GlobalPatterns-based distribution (Figure 4).

8. A filtering step was applied to remove rare taxa from the simulated datasets, which may have skewed the shape of the community distributions

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 (35–37). However, my previous work has shown that rare sequences from a poorly sequenced sample often appear in more deeply sequenced samples suggesting that they are not necessarily artifacts (4). Furthermore, removing rare sequences alters the structure of communities and has undesirable effects on downstream analyses. 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 calculated the the difference in accuracy between replicates for each normalization method and distance calculator across effect sizes for a median sequencing depth of 10,000 (Figure S8). The median difference in accuracies (i.e. clustering accuracies with filtered data minus those without filtering) did not deviate meaningfully from zero. However, the 95% confidence intervals were most pronounced at large effect sizes when using raw counts, DESeq Variance Stabilization and Upper-Quartile Log-Fold Change and at smaller effect sizes when using rarefaction and relative abundances. Given the large variation caused by removing rare taxa and my previous work (4), OTU filtering should not be performed in microbiome analyses. Considering the minimal effect that removing rare OTUs had on the median difference in clustering accuracy in the current simulation framework, I have used the filtered datasets throughout the current study.

9. There was no accounting for differences in performance when library sizes are confounded with treatment group

I and others have observed that not using rarefaction can lead to falsely detecting differences between communities when sequencing effort is confounded with the treatment group (25, 38). Previous analyses showed that in these situations 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. 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 (e.g., 39). To determine whether this result was replicated with the WNWN simulation framework, I created a sequencing depth distribution where sequencing depth was fully confounded with the treatment group using both the GlobalPatterns and Log-scaled sequence distributions. To confound the sequencing depth, sequencing depths from one treatment group were drawn from below the median number of sequences of the sequencing distribution and those for the second treatment group were from above the median. To assess the risk of falsely detecting clusters, I compared the clustering accuracies using an effect size of 1.00 using both the confounded and randomized sequencing distributions (see rows 1 and 2 of Figure 6). 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 were two groups of 40 samples that did 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 false detection risk did not vary by method when the sequencing depth was randomized across treatment groups. Yet, when the sequencing depth was confounded with treatment group, rarefaction was the most consistent normalization method for controlling the generation of spurious clusters. At larger effect sizes the ability to correctly identify two clusters increased when the sequencing depth was confounded with treatment group (Figure 6). At the effect size of 1.15, the rarefied data generated the highest accuracy clusters regardless of whether the sequencing depths were confounded. Although the level of confounding in this simulation was extreme, it highlights the ability of rarefaction to control the false detection rate and the ability to correctly detect clusters relative to the other normalization methods.

10. Clustering accuracy was used rather than the more direct and frequently applied comparisons of beta diversity using permutation tests

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 (28, 40, 41). 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 1.00, it was impossible for all 80 samples to be assigned to a single cluster. A more commonly used approach for analyzing distance matrices is to use a non-parametric analysis of variance test (i.e., AMOVA, PERMANOVA, NP-ANOVA)(42). 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 (43). When sequencing depths were randomly distributed across the two treatment groups, the Type I error did not meaningfully deviate from the expected 5% (Figure 7). However, when sequencing depths were confounded with treatment group, 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 7). 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 richness, which is the traditional application of rarefaction

Rarefaction was originally proposed as a method for controlling uneven sequencing effort when comparing community richness values (5, 6). Thus it was surprising that WNWN did not consider the effect of the proposed normalizations on richness and 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 Variance Stabilization 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 whether the alpha diversity was significantly different between the treatment groups 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. For these analyses, I used the GlobalPatterns (Figure 8) and log-distributed (Figure S9) data with the random and confounded distribution of sequencing depths. Similar to the results in points 9 and 10, with the exception of rarefaction, the simulations using a confounded sequence depth distribution resulted in all of the replicates having a significant test. 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 0.91 and 1.00, respectively. The power then decreased with increasing sequencing effort to 0.57 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 and average of 1,559.65 OTUs and the ocean had an average of 1,335.00 OTUs; they shared an average of 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 significant 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 one faces when comparing communities using metrics based on presence-absence data like richness and Jaccard and Unweighted UniFrac distances.

Discussion

The conclusions from WNWN have had a lasting impact on how researchers analyze microbiome sequence data. As I have demonstrated using WNWN's simulation framework, 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 remove 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 the word “rarefaction”. As I have demonstrated there were numerous choices throughout the original study that resulted in rarefying/subsampling looking worse than it truly is. In fact, when the data of the current study 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 sequencing effort when analyzing alpha and beta diversity metrics. Although beyond the scope of the current study, the same principles would also support using rarefaction to control for uneven sequencing effort when calculating alpha and beta diversity metrics in shotgun metagenomic sequencing analyses.

It is worth commenting on WNWN'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. While this assumption might be valid when comparing host gene expression in 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 absence of the OTU. In fact, zeroes are less common in gene expression analyses. But because microbiome studies have so many zeroes, it is necessary to add a pseudo-count to all OTUs for both the edgeR and DESeq-based normalizations (L443 and L487, respectively). In WNWN, a pseudo-count of 1 was used. However, this value is arbitrary and the results can vary based on the choice of pseudo-count and the patchiness of the communities being analyzed. Since WNWN was published, compositional approaches have been proposed to account for uneven sequencing and to provide improved interpretability (14, 16–18, 44, 45). However, these methods also often require the use of pseudo-counts. 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 of selecting one. Within the ecology literature, Euclidean distances are widely avoided because joint absence is weighted the same as joint presence of taxa (30). As discussed in point 11, distances that are based on community membership (i.e., the presence or absence 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 distances 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 distances. In the current analysis, the phylogenetic Weighted UniFrac distance performed comparably to Bray-Cutis distances for clustering or differentiating between communities with `adonis2`. Regardless of the distance calculation employed, they all performed best when using rarefaction relative to the other normalization methods.

The WNWN 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 actual microbiome studies the selection of a sequencing 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-addressing this set of simulations is beyond the scope of the current analysis and others have already contributed critiques (25). Many of the same concerns addressed here would also apply. Perhaps most important is the fact that if the relative abundance of several OTUs increase, then the relative abundance of all other OTUs would necessarily decrease because the data are compositional (45). 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 independent of all other OTUs. This is an important limitation of tests of methods attempting to detect differentially abundant OTUs. Alternatively, one could increase the abundances of several OTUs while decreasing the abundances of other OTUs without changing the overall total. Regardless, a form of rarefaction could still be employed for detecting differential abundance. One could subsample the data, perform the statistical test, and identify the differentially abundant OTUs. In my experience the largest differences between subsamples are for low relative abundances OTUs, which are unlikely to be statistically significant. This process could be repeated a small number of times to confirm that the significance of OTUs are robust to subsamplings.

In a parallel set of analyses I have used alternative simulation and evaluation strategies to look more closely at rarefaction and its alternatives (38) and the practice of filtering low abundance sequences (4). The results of those studies are similar to this analysis. These studies demonstrate that it is actually sequence and OTU filtering that “requires the omission of available valid data” and that rarefaction is the only available method for controlling the effects of uneven sequencing. Far from being inadmissible, rarefaction of unfiltered datasets yields the most robust results when comparing communities based on alpha and beta diversity metrics.

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_XXXXX_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 the repository). This manuscript was written as an R 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 the repository).

Reproducing WNWN Protocol S1

The compressed directory published with WNWN as Protocol S1 is available on the *PLOS Computational Biology* website with WNWN at https://doi.org/10.1371/journal.pcbi.1003531.s001. To render the R markdown files to HTML files using software and packages as similar as possible to that of WNWN, I created an environment (see `workflow/envs/nr-s1.yml` in this study's 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 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 environment. Both changes were necessary to better organize the project and had no substantial bearing on the content of the rendered HTML file.

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 generating 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 effect sizes used in the WNWN study (1.00 [i.e., no difference], 1.15, 1.25, 1.50, 1.75, 2.00, 2.50, and 3.50). Second, the number of sequences in each of the 80 samples was either randomly selected from the 26 library sizes contained within the GlobalPatterns dataset or from a log distribution of 80 sequencing depths evenly distributed between the most shallow (58,688 sequences) and deeply (2,357,181 sequences) sequenced samples in the GlobalPatterns dataset (Figure S6). 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, or 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 effect sizes x 2 sequencing depth models x 5 sequencing depths x 2 sequencing depth assignment models x 100 replicates).

Generation of distances between communities

Next, each simulation was further processed by filtering rare OTUs, to normalize uneven sequencing 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 unfiltered. As was done in the 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) DESeq Variance Stabilization, (4) Upper-Quartile Log-Fold Change normalization using edgeR, (5) a single subsampling to a common number of sequences, (6) rarefaction with 100 randomizations to a common number of sequences. Although the data were not presented, the WNWN code 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 WNWN 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 normalizations and distances to visualize was determined by the WNWN 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 methods 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` R package, K-means clustering with the `kmeans` function from the `stats` base R package, and hierarchical clustering with the `hclust` and `cutree` functions from the `stats` base R package. The accuracy of the clustering was measured as the fraction of the 80 samples assigned to the correct cluster. As in WNWN, if samples were removed because the number of sequences in them fell below the threshold for subsampling or rarefaction, then the accuracy could be below 50%. For the effect size of 1.00, 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 samples (i.e., an accuracy of 0.51). 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` Conda 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 Type I error rate for a condition was the fraction of significant tests when the effect size was 1.00. The power was the fraction of significant tests at other effect sizes.

Analysis of richness and diversity between communities

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 formula (30). For the rarefaction data, each of the 80 samples in the 100 subsampled datasets were used to calculate the richness and diversity. The mean across the 100 subsamplings was used as the value for the rarefaction data across each of the 80 samples. 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 R package. Similar to the analysis using `adonis2`, 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 Type I error rate for a condition was the fraction of significant tests when the effect size was 1.00. The power was the fraction of significant tests at other effect sizes.

Acknowledgements

I am grateful to Dr. Paul McMurdie and Dr. Susan Holmes for publishing the source code that they used to conduct their analysis as R markdown files in the supplement to WNWN. It would not have been possible for me to replicate and build upon their simulations without their code.

References

1. **Sogin ML**, **Morrison HG**, **Huber JA**, **Welch DM**, **Huse SM**, **Neal PR**, **Arrieta JM**, **Herndl GJ**. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proceedings of the National Academy of Sciences **103**:12115–12120. doi:10.1073/pnas.0605127103.

2. **Caporaso JG**, **Lauber CL**, **Walters WA**, **Berg-Lyons D**, **Lozupone CA**, **Turnbaugh PJ**, **Fierer N**, **Knight R**. 2010. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences **108**:4516–4522. doi:10.1073/pnas.1000080107.

3. **Kozich JJ**, **Westcott SL**, **Baxter NT**, **Highlander SK**, **Schloss PD**. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq illumina sequencing platform. Applied and Environmental Microbiology **79**:5112–5120. doi:10.1128/aem.01043-13.

4. **Schloss PD**. 2020. Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data. bioRxiv. doi:10.1101/2020.12.11.422279.

5. **Sanders HL**. 1968. Marine benthic diversity: A comparative study. The American Naturalist **102**:243–282. doi:10.1086/282541.

6. **Hurlbert SH**. 1971. The nonconcept of species diversity: A critique and alternative parameters. Ecology **52**:577–586. doi:10.2307/1934145.

7. **Anders S**, **Huber W**. 2010. Differential expression analysis for sequence count data. Genome Biology **11**:R106. doi:10.1186/gb-2010-11-10-r106.

8. **Beule L**, **Karlovsky P**. 2020. Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): Application to microbial communities. PeerJ **8**:e9593. doi:10.7717/peerj.9593.

9. **Chao A**, **Shen T-J**. 2003. Environmental and Ecological Statistics **10**:429–443. doi:10.1023/a:1026096204727.

10. **Chao A**, **Chiu C-H**. 2016. Species richness: Estimation and comparison, p. 1–26. *In* Wiley StatsRef: Statistics reference online. John Wiley & Sons, Ltd.

11. **Costea PI**, **Zeller G**, **Sunagawa S**, **Bork P**. 2014. A fair comparison. Nature Methods **11**:359–359. doi:10.1038/nmeth.2897.

12. **Hughes JB**, **Hellmann JJ**, **Ricketts TH**, **Bohannan BJM**. 2001. Counting the uncountable: Statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology **67**:4399–4406. doi:10.1128/aem.67.10.4399-4406.2001.

13. **Hughes JB**, **Hellmann JJ**. 2005. The application of rarefaction techniques to molecular inventories of microbial diversity, p. 292–308. *In* Methods in enzymology. Elsevier.

14. **Lin H**, **Peddada SD**. 2020. Analysis of microbial compositions: A review of normalization and differential abundance analysis. npj Biofilms and Microbiomes **6**:60. doi:10.1038/s41522-020-00160-w.

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

16. **Martino C**, **Morton JT**, **Marotz CA**, **Thompson LR**, **Tripathi A**, **Knight R**, **Zengler K**. 2019. A novel sparse compositional technique reveals microbial perturbations. mSystems **4**:00016–19. doi:10.1128/msystems.00016-19.

17. **Paulson JN**, **Stine OC**, **Bravo HC**, **Pop M**. 2013. Differential abundance analysis for microbial marker-gene surveys. Nature Methods **10**:1200–1202. doi:10.1038/nmeth.2658.

18. **Quinn TP**, **Erb I**, **Gloor G**, **Notredame C**, **Richardson MF**, **Crowley TM**. 2019. A field guide for the compositional analysis of any-omics data. GigaScience **8**:giz107. doi:10.1093/gigascience/giz107.

19. **Robinson MD**, **McCarthy DJ**, **Smyth GK**. 2009. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**:139–140. doi:10.1093/bioinformatics/btp616.

20. **Beest DE te**, **Nijhuis EH**, **Möhlmann TWR**, **Braak CJF**. 2021. Log-ratio analysis of microbiome data with many zeroes is library size dependent. Molecular Ecology Resources **21**:1866–1874. doi:10.1111/1755-0998.13391.

21. **Willis AD**. 2019. Rarefaction, alpha diversity, and statistics. Frontiers in Microbiology **10**:2407. doi:10.3389/fmicb.2019.02407.

22. **Willis A**, **Bunge J**. 2015. Estimating diversity via frequency ratios. Biometrics **71**:1042–1049. doi:10.1111/biom.12332.

23. **McKnight DT**, **Huerlimann R**, **Bower DS**, **Schwarzkopf L**, **Alford RA**, **Zenger KR**. 2018. Methods for normalizing microbiome data: An ecological perspective. Methods in Ecology and Evolution **10**:389–400. doi:10.1111/2041-210x.13115.

24. **McMurdie PJ**, **Holmes S**. 2014. Waste not, want not: Why rarefying microbiome data is inadmissible. PLoS Computational Biology **10**:e1003531. doi:10.1371/journal.pcbi.1003531.

25. **Weiss S**, **Xu ZZ**, **Peddada S**, **Amir A**, **Bittinger K**, **Gonzalez A**, **Lozupone C**, **Zaneveld JR**, **Vázquez-Baeza Y**, **Birmingham A**, **Hyde ER**, **Knight R**. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome **5**:27. doi:10.1186/s40168-017-0237-y.

26. **Gotelli NJ**, **Colwell RK**. 2001. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4**:379–391. doi:10.1046/j.1461-0248.2001.00230.x.

27. **Navas-Molina JA**, **Peralta-Sánchez JM**, **González A**, **McMurdie PJ**, **Vázquez-Baeza Y**, **Xu Z**, **Ursell LK**, **Lauber C**, **Zhou H**, **Song SJ**, **Huntley J**, **Ackermann GL**, **Berg-Lyons D**, **Holmes S**, **Caporaso JG**, **Knight R**. 2013. Advancing our understanding of the human microbiome using QIIME, p. 371–444. *In* Methods in enzymology. Elsevier.

28. **Koren O**, **Knights D**, **Gonzalez A**, **Waldron L**, **Segata N**, **Knight R**, **Huttenhower C**, **Ley RE**. 2013. A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. PLoS Computational Biology **9**:e1002863. doi:10.1371/journal.pcbi.1002863.

29. **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. doi:10.1038/s41598-021-01636-1.

30. **Legendre P**, **Legendre L**. 2012. Numerical ecology. Elsevier Science.

31. **Lozupone C**, **Knight R**. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology **71**:8228–8235. doi:10.1128/aem.71.12.8228-8235.2005.

32. **Lozupone CA**, **Hamady M**, **Kelley ST**, **Knight R**. 2007. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and Environmental Microbiology **73**:1576–1585. doi:10.1128/aem.01996-06.

33. **McMurdie PJ**, **Holmes S**. 2013. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE **8**:e61217. doi:10.1371/journal.pone.0061217.

34. **Witten DM**. 2011. Classification and clustering of sequencing data using a poisson model. The Annals of Applied Statistics **5**:2493–2518. doi:10.1214/11-aoas493.

35. **Amir A**, **McDonald D**, **Navas-Molina JA**, **Kopylova E**, **Morton JT**, **Xu ZZ**, **Kightley EP**, **Thompson LR**, **Hyde ER**, **Gonzalez A**, **Knight R**. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems **2**:00191–16. doi:10.1128/msystems.00191-16.

36. **Bokulich NA**, **Subramanian S**, **Faith JJ**, **Gevers D**, **Gordon JI**, **Knight R**, **Mills DA**, **Caporaso JG**. 2013. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. Nature Methods **10**:57–59. doi:10.1038/nmeth.2276.

37. **Edgar RC**. 2016. UNOISE2: Improved error-correction for illumina 16S and ITS amplicon sequencing. bioRxiv. doi:10.1101/081257.

38. **Schloss PD**. 2023. Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. bioRxiv. doi:10.1101/2023.06.23.546313.

39. **Morris A**, **Beck JM**, **Schloss PD**, **Campbell TB**, **Crothers K**, **Curtis JL**, **Flores SC**, **Fontenot AP**, **Ghedin E**, **Huang L**, **Jablonski K**, **Kleerup E**, **Lynch SV**, **Sodergren E**, **Twigg H**, **Young VB**, **Bassis CM**, **Venkataraman A**, **Schmidt TM**, **Weinstock GM**. 2013. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. American Journal of Respiratory and Critical Care Medicine **187**:1067–1075. doi:10.1164/rccm.201210-1913oc.

40. **Knights D**, **Ward TL**, **McKinlay CE**, **Miller H**, **Gonzalez A**, **McDonald D**, **Knight R**. 2014. Rethinking “enterotypes”. Cell Host & Microbe **16**:433–437. doi:10.1016/j.chom.2014.09.013.

41. **Costea PI**, **Hildebrand F**, **Arumugam M**, **Bäckhed F**, **Blaser MJ**, **Bushman FD**, **Vos WM de**, **Ehrlich SD**, **Fraser CM**, **Hattori M**, **Huttenhower C**, **Jeffery IB**, **Knights D**, **Lewis JD**, **Ley RE**, **Ochman H**, **O'Toole PW**, **Quince C**, **Relman DA**, **Shanahan F**, **Sunagawa S**, **Wang J**, **Weinstock GM**, **Wu GD**, **Zeller G**, **Zhao L**, **Raes J**, **Knight R**, **Bork P**. 2017. Enterotypes in the landscape of gut microbial community composition. Nature Microbiology **3**:8–16. doi:10.1038/s41564-017-0072-8.

42. **Anderson MJ**. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology **26**:32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x.

43. **Dixon P**. 2003. VEGAN, a package of r functions for community ecology. Journal of Vegetation Science **14**:927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.

44. **Palarea-Albaladejo J**, **Martin-Fernandez J**. 2015. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems **143**:85–96. doi:10.1016/j.chemolab.2015.02.019.

45. **Gloor GB**, **Macklaim JM**, **Pawlowsky-Glahn V**, **Egozcue JJ**. 2017. Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology **8**:2224. doi:10.3389/fmicb.2017.02224.

Figures

Figure 1. Rarefaction resulted in larger and less variable clustering accuracies. With the exception of Unweighted UniFrac distances, the improved performance by rarefaction was observed at smaller effect sizes. In the first row of panels larger values mean that the accuracies by rarefaction were better than those of subsampling. In the second row of samples, larger values mean that interquartile range (IQR) for rarefaction was larger than that of subsampling.

Figure 2. K-means clustering was consistently as good or better than PAM or hierarchical clustering when comparing rarefaction to other normalization methods. Each point represents the percentage of 100 simulations where that clustering method performed as well or better than the other methods for that normalization procedure.

Figure 3. When the median sequencing depth was 2,000 sequences or more, rarefaction of the entire dataset performed better than removing the smallest 15% of samples when using K-means clustering. This figure is analogous to Figure S4 except that K-means clustering was used instead of PAM.

Figure 4. K-means clustering of distances calculated with rarefaction were as good or better than any other normalization method. This figure is analogous to Figure S3 except that K-means clustering was used instead of PAM, rarefaction on the full dataset was used instead of subsampling to the size of the sample at the 15th percentile, and DESeq Variance Stabilization normalized OTU counts were only used with Euclidean distances.

Figure 5. Clustering accuracies that used rarefaction were as good or better than the other normalization procedures when a log-scaled distribution of sequencing depths. This figure is analogous to Figure 4 except that the sequencing depths for each of the 80 samples in each simulation were drawn without replacement from a log-scaled distribution rather than from the GlobalPatterns sequencing depths.

Figure 6. Rarefaction was consistently as good or better than all other normalization methods at assigning samples to the correct treatment group regardless of whether sequencing depth was confounded by treatment group. Because the clustering algorithms forced samples into one of two groups, the expected accuracy with an effect size of 1.00 was 0.51. With an effect size of 1.15, the expected accuracy was 1.00. Each point represents the median of 100 replicates and the error bars represent the observed 95% confidence interval. Data are shown for a median sequencing depth ($\tilde{N}_L$) of 10,000 sequences when individual sequencing depths were sampled with replacement from the

GlobalPatterns dataset or without replacement from the log-scaled distribution.

Figure 7. Rarefaction was consistently as good or better than all other normalization methods at controlling for Type I error and maximizing power to detect differences in treatment group using adonis2 regardless of whether sequencing depth was confounded by treatment group. Type I errors were assessed as the fraction of 100 simulations that yielded a significant P-value (i.e., less than or equal to 0.05) at an effect size of 1.00. Power was assessed as the fraction of 100 simulations that yielded a significant P-Value at an effect size of 1.15. Data are shown for a median sequencing depth ($\tilde{N}_L$) of 10,000 sequences when individual sequencing depths were sampled with replacement from the GlobalPatterns dataset or without replacement from the log-scaled distribution.

Figure 8. Rarefaction was consistently as good or better than all other normalization methods at controlling for Type I error and maximizing power to detect differences in treatment group using alpha-diversity metrics regardless of whether sequencing depth was confounded by treatment group when using sequencing depths drawn from the GlobalPatterns datasets. Statistical comparisons of OTU richness and Shannon diversity were performed using the non-parametric Wilcoxon two-sampled test. Type I errors were assessed as the fraction of 100 simulations that yielded a significant P-value (i.e., less than or equal to 0.05) at an effect size of 1.00. Power was assessed as the fraction of 100 simulations that yielded a significant P-Value at an effect size of 1.15. Data are shown for when the case when individual sequencing depths were sampled with replacement from the GlobalPatterns dataset.

Figure S1. Re-running the R markdown files provided in Protocol S1 of WNWN qualitatively reproduced Figure 4 from WNWN.

Figure S2. Re-running the R markdown files provided in Protocol S1 of WNWN qualitatively reproduced Figure 5 from WNWN.

Figure S3. Successful reimplementation and expansion of analysis presented in Figure 4 from WNWN in a Snakemake pipeline. The reimplemented workflow largely borrowed from the original `simulation-cluster-accuracy-server.Rmd` R markdown file provided in WNWN's Protocol S1. The most notable differences include the use of 100 rather than 5 randomizations and the addition of the median sequencing depth ($\tilde{N}_L$) of 50,000. The plotting symbols indicate the median of 100 randomizations and the error bars represent the observed 95% confidence interval. Simulations run at the same effect size are dodged to better reveal overlapping data. The sequencing depths were drawn from the GlobalPatterns dataset and sequences were clustered using PAM.

Figure S4. Successful reimplementation and expansion of analysis presented in Figure 5 from WNWN in a Snakemake pipeline. The reimplemented workflow largely borrowed from the original `simulation-cluster-accuracy-server.Rmd` R markdown file provided in WNWN's Protocol S1. The most notable differences include the use of 100 rather than 5 randomizations and the addition of the median sequencing depth ($\tilde{N}_L$) of 50,000. The plotting symbols indicate the median of 100 randomizations and the error bars represent the observed 95% confidence interval. Simulations run at the same effect size are dodged to better reveal overlapping data. A light gray line is shown to indicate the best possible accuracy for each library size minimum quantile value. The sequencing depths were drawn from the GlobalPatterns dataset and sequences were clustered using PAM.

Figure S5. DESeq Variance Stabilization of OTU counts resulted in negative values that were used to calculate Bray-Curtis and Weighted UniFrac distances. The median number of negative OTU counts that had a negative OTU count following normalization increased with the effect size and decreased as $\tilde{N}_L$ increased (first row). The error bars indicate the observe 95% confidence interval. The fraction simulated datasets that had a negative OTU count following normalization also increased with effect size, but increased as $\tilde{N}_L$ increased (second row). For each effect size there were 100 replicate datasets.

Figure S6. Comparison of the normally distributed sequencing depths from the GlobalPatterns dataset and a log-scaled distribution of sequencing depths. The log-scaled distribution was generated so that each sample in a simulation could have a unique number of sequences and to simulate

the skew right distribution commonly seen in microbiome studies.

Figure S7. Rarefaction with all samples yielded clustering accuracies that were as good or better than removing the smallest 15% of samples across distance calculation methods. This figure is analogous to Figure 8 except that the sequencing depths for each of the 80 samples in each simulation were drawn without replacement from a log-scaled distribution rather than from the GlobalPatterns sequencing depths.

Figure S8. Normalization and distance calculation methods vary in their sensitivity to removal of rare OTUs. Larger values indicate that the clustering accuracy from filtered datasets were larger than those from non-filtered datasets. The median of 100 randomizations did not meaningfully vary from 0.0, but the observed 95% confidence interval varied considerably. Data are shown for a median sequencing depth ($\tilde{N}_L$) of 10,000 sequences when individual sequencing depths were sampled with replacement from the GlobalPatterns dataset or without replacement from the log-scaled distribution.

Figure S9. Rarefaction was consistently as good or better than all other normalization methods at controlling for Type I error and maximizing power to detect differences in treatment group using alpha-diversity metrics regardless of whether sequencing depth was confounded by treatment group when using sequencing depths drawn from a lognormal distribution. Statistical comparisons of OTU richness and Shannon diversity were performed using the non-parametric Wilcoxon two-sampled test. Type I errors were assessed as the fraction of 100 simulations that yielded a significant P-value (i.e., less than or equal to 0.05) at an effect size of 1.00. Power was assessed as the fraction of 100 simulations that yielded a significant P-Value at an effect size of 1.15. Data are shown for when the case when individual sequencing depths were sampled without replacement from the log-scaled distribution.