```
## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3
```

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

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

Patrick D. Schloss†

† To whom corresponsdence should be addressed:

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Research article

Abstract

Introduction

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

In 2014 Paul McMurdie and Susan Holmes published their "Wast not, want not: why rarefying microbiome data is inadmissible" (WNWN) in PLOS Computational Biology [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 2,300 times as of January 2023. Anecdotely, I have received correspondence from researchers over the past 10 years asking how to address critiques from reviewers who criticize my correspondents' analysis for rarefying (e.g., see this Twitter thread). I have also received these types of comments from reviewers, specifically in regards to a preprint that I posted in 20XX in regards to my critique of the practice of removing rare taxa from analyses [XXXXX]. In the process of responding to these critiques 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 purported to rarefy the data only performed one subsampling of the data (Lines 404 through 416 of `simulation-cluster-accuracy/simulation-cluste`
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. 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.

Adding to the confusion is that the papers cited in the first sentence of the quote I WNWN included above either do not use the words “rarefy” or “rarefying” or use them interchangably with “rarefaction”. In hindsight, as shown in the quoted text, McMurdie and Holmes do emphasize the distinction between rarefying and rarefaction. However, because they seem to have coined a new meaning for rarefying, they seem to have only added to the confusion by using the generally used verb form of rarefaction. Further confusion comes from the author’s admonition in the final sentence that some 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 rarefaction with a single randomization.

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

1. Select a minimum library size, $N_{L,m}$. 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. It is important to note that this procedure generates substantially different results to those obtained without accounting for uneven sampling effort or using relative abundances, normalized counts, compositional data transformatins, variance stabilization procedures, or estimation techniques. I have explored the differences in results obtained using the diversity of approaches for controlling for uneven sampling effort; rarefaction, as described here, outperforms the other approaches [XXXXX].

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. 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 samples to one of two clusters using simulated data used to generate 40 samples from each of two distributions. A variety of approaches were used to calculate distances between the samples those distances were used as input to partitioning around mediods (PAM) clustering with two

groups. The accuracy of the cluster assignment was used as the metric to assess performance. This analysis was performed in the `simulation-cluster-accuracy/simulation-cluster-accuracy-server.Rmd` R-flavored markdown file that was published as Protocol S1 in the original paper. Line numbers from this file will be refernced with an "L" as a prefix.

The two distributions were generated using human fecal and ocean data originally take from the GlobalPaterns dataset (L129) [XXXXX]. To generate a fecal and ocean template 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 2000 most common and abundant OTUs and pooled the abundances of the fecal and ocean samples separately to create two templates (L144, L159-160, L197-198).

Next, the fecal and ocean templates were mixed in 8 different fractions to generate two community types that differend by varying effect sizes (1, 1.15, 1.25, 1.5, 1.75, 2, 2.5, and 3.5; L170-195, L220). To simulate the variation in sequencing depth across the 80 samples, they normalized he number of sequences from each of the `gp_n_samples` samples in the GlobalPatterns dataset so that the median number of sequences (N_L) for the GlobalPatterns had 1,000, 2,000, 5,000, or 10,000 sequences (L324-325). They then randomly sampled the `gp_n_samples` normalized sequencing depths to generate 80 sampling depths. From each community type, they simulated 40 samples by sampling to the desired number of reads (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).

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, which caused results to be sensitive to ran-

dom number generator seed

2. The average sizes of the libraries were small by modern standards
3. DeSeq-based variance stabilization was used with inappropriate distance calculation methods
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 poorly chosen
8. A filtering step was applied to remove rare taxa from the simulated datasets, which could distort the shape of the communities
9. No test of effect of transformation methods to account for effects of uneven sample size between treatment groups
10. Clustering accuracy was used rather than direct comparisons beta diversity
11. No consideration of effects of transformations 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 and my own. 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 get the document to render. 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 their 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. Regardless of the differences, the results are qualitatively similar.

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 DEeqVS transformed 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 are unlikely to be symmetric around a mean at larger 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 as above. This was necessary since the number of simulated conditions because of this change increased 20-fold. 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 S4 (**pam_subsample15_fig_4.pdf**) and S5(**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 average depth of sampling value of 50,000 sequences with the four average sequencing sampling depths as WNWN (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 inappropriate distance calculation methods

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 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 a warning message 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 transformed 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 transformation with the Euclidean distance.

4. A single subsampling of each dataset was evaluated rather than using rarefaction

As noted above, the original jargon that was used in WNWN was confusing to many who conflated rarefying with rarefaction. One problem with a single subsampling of a community is that it is unlikely to obtain a representative sampling of each community. A more robust analysis would have used rarefaction rather than a single subsampling of the data 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", 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 distances from a single subsampling to rarefaction using the distance calculations shown in the original Figure 4. This analysis revealed two benefits of rarefaction. First, the median distances generated by rarefaction was always as large or larger than those from a single subsample (Figure **subsample_rarefaction_compare.pdf**). In general, 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 intraquartile range by subsampling 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, close inspection of the rendered figures suggests that PAM may not have been the optimal choice in all situations. Rather, K-means clustering may have been prefered. 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 when 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

In WNWN, the authors quantified the tradeoff between sampling effort, the number of samples removed below the threshold, and clustering accuracy (original Figure 5, my Figure S5). Although the optimal sampling effort varied by distance metric, transformation method, and sampling effort, 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% (Figure S2). Because the true result would not known *a priori* in microbiome studies, 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. [**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.**] 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 sam-

ples (**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 for with the Poisson distance (Figure **kmeans_rarefaction15_fig_4.pdf**). It is worth noting that at larger effect sizes, K-means clustering did not perform as well for some combinations of normalization methods and distance calculations (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 poorly chosen

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 1,085,256.85 sequences and a median of 1,106,849.00 sequences per dataset. The datasets ranged in sequencing depth between 58,688.00 and 2,357,181.00 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 transformations and distance calculations are 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.

I created a new set of simulations that would focus on solely on the effect of the shape of the distribution on the results to reassess the WNWN results in the context of a more typical distribution of sample sizes. 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 372,040.00 and the mean was 629,825.00. When I regenerated the original Figures 4 and 5 using the log-distributed sequencing effort distribution, the differences in transformation methods were more apparent (Figure ***fig_4_kmeans_rarefaction00_log.pdf***). For each of the distance calculators, rarefying the data 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 number of samples 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 biological samples [XXXXXX].

8. A filtering step was applied to remove rare taxa from the simulated datasets, which could distort the shape of the communities

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 and analyzing their simulated communities. 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 appear 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,

relative abundance had the narrowest range at small effect sizes; at larger effect sizes, rarefaction had the narrowest range. Again, these trends appear 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 test of effect of transformation methods to account for effects of uneven sample size between treatment groups

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**). For the case when the effect size was 1.0, 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 small effect size of 1.15, with the exception of the Poisson distance data, which generally clustered performed poorly, the rarefied data generated the highest accuracy clusters regardless of whether the data were skewed. Although the skewed simulation is extreme, it highlights the ability of rarefaction to control Type I errors while maintaining high power.

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

The analysis of the GlobalPatterns-based simulations in WNWN used clustering accuracy to assess the effect of different normalization procedures, distance calculations, and clustering methods. There has been controversy over the meaning of clustering samples (i.e., enterotypes) 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 demonstrated in point 5, an additional problem wiht clustering is the sensitivity to the chosen method. A more common analysis of distance matrices is to use a non-parametric analysis of variance test of the various distance matrices (i.e., AMOVA, PERMANOVA, adonis)[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 transformation 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 transformation 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 the tradeoff between Type I error and power, rarefaction performed the best among the different transformations within this simulation framework.

11. No consideration of effects of transformations on alpha diversity metrics

Rarefaction was originally proposed as a method for controlling uneven sampling effort when comparing community richness values. Thus it was surprising that WNWN did not consider the effect of the proposed transformations on richness or other alpha-diversity metrics such as Shannon diversity. Therefore, for each of the transformations, I compared the richness and diversity of the two treatment groups. The DeSeq transformed data were not included because the transformation produced negative values which were not compatible with calculations of richness or Shannon diversity. Also, data from the Upper Quartile Log Fold Change transformation were not used for richness calculations since the transformation 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.

As shown above, I compared the risk of committing Type I errors and the power to detect differences by the different transformations (Figure **alpha_compare.pdf**). For these analyses, I used the GlobalPatterns data with the random and skewed distribution of samples. As shown earlier, the simulations using a skewed distribution resulted in all of the replicates having a significant test result when using any of the tranformations except rarefaction. The power to detect differences in richness and diversity at effect sizes of 1.15 and greater with rarefaction was at least as high or greater than any of the other transformations when samples were randomly distributed between the two treatment groups. Again, these results demonstrate that for alpha diversity measurements, rarefaction is the preferred approach of controlling for uneven sampling effort.

Discussion

- Gratidue that code was published with paper. This made it straight forward to notice that only a single subsampling step was performed for each random seed and that only 3 random seeds were used.
- Differential abundance issues
- How we actually pick a threshold
- UniFracs are no longer viable methods since trees are too challenging to construct *de novo* with modern library sizes
- Estimating richness
- Saturate richness at large number of sequences
  - 952 of 2000 OTUs overlap between Ocean and Feces

Acknowledgements

References

Figures