Thank you for submitting your manuscript to mSphere. We have completed our review and I am pleased to inform you that, in principle, we expect to accept it for publication in mSphere. However, acceptance will not be final until you have adequately addressed the reviewer comments. I am impressed by the quality of the reviews and hope that you find them useful. I share their enthusiasm for the value of the paper to the research community and believe it will be widely read.

> I am grateful for the thoughtful feedback from both reviewers. Given the lengthy manuscript, I truly appreciate their efforts to consider the overall message. I have done my best to address their comments and am confident the manuscript is significantly improved because of their feedback. Since posting the manuscript to bioRxiv and submitting it to *mSphere*, I have received numerous emails from people thanking me for this contribution and expressing their eagerness to see the work published. In my responses to the reviewers' comments when I refer to line numbers these are from the marked up version of the manuscript.

I'm a bit concerned about the number of figures (17). I know that we do not have an upper limit for Research Articles, but it still seems like a lot for the main text. We could certainly start a more philosophical discussion of whether it matters in these days of fully electronic publications. :-)

> Well, I tried :) In the revised manuscript I have kept what I consider to be the most important 8 figures in the main body of the manuscript and moved the other 9 figures in to the supplement.

Reviewer #1 (Comments for the Author):

General Comments: This paper takes issue with a famous paper, "WNWN", that argued against rarefying data to address uneven library sizes. It starts by discussing that the terms rarefy and rarefaction are often confused, and are alternately used to refer to a single subsampling or to multiple subsampling. WNWN used single subsampling in its analysis that showed rarefaction was inferior to normalizing to percent reads. This paper sets out to prove that multiple subsampling (and then averaging the analysis results) is actually superior to both single subsampling and normalizing to percent reads. The paper identifies 11 issues with the WNWN paper (#4 of which is the single subsampling), and repeats the first simulation in that paper with adjusted metrics. Overall, the manuscript is thorough and well written, and clarifies an important point of common confusion. I think it will be valuable to the field as this alternate, multiple subsampling method has been unfairly dismissed largely due to confusion over terminology.

I am convinced by this paper that multiple subsampling is the best method to account for uneven library sizes when calculating diversity metrics. However, my main concern is that there is no acknowledgement that most modern manuscripts will do much more with 16S data than diversity analyses, and rarefaction is often not computationally feasible because it would require all analyses to be repeated 100-1000 times and the results averaged. Therefore, the choice addressed by WNWN is often the practical one- either single-subsample or normalize- so stating that "their claims are not supported" (line 470) feels like a reach, especially since it is based largely on an argument about semantics. WNWN showed that single supsampling should not be used, this is still supported in this paper; this paper also argues that they called single subsampling the wrong thing. What this paper shows is that single subsampling (and normalizing) does not perform as well as rarefaction for a very specific analysis to which it can be easily applied (diversity metrics). To be truly applicable to a broad audience performing myriad analyses, this paper must address this narrow focus directly. It seems to me that normalizing is likely still the best practical approach for many non-diversity analyses, yet this paper argues emphatically that rarefaction is "superior." The limited scope of applications of this "superior" method should be discussed, as well as an acknowledgement that the WNWN paper was focused on broader applications too.

> I'm afraid that the reviewer and I will have to disagree on this point. Rarefaction of alpha and beta diversity metrics is practical with modern datasets and is found in nearly every microbiome analysis. The algorithm for rarefaction is parallelized in mothur for quicker calculation and can easily be parallelized for tests of differential abundance. WNWN showed a single subsampling should not be used by making the method look as bad as possible by deducting 15% points from its accuracy, using PAM instead of K-means, or using alpha and beta diversity metrics instead of clustering. Rarefaction is simply superior and because of the framing of the jargon and analyses in WNWN people have been using substandard methods for the past decade.

Specific Comments:

57- link is dead (and maybe so is twitter?) Instead of linking to the preprints and tweets, it'd flow more nicely within the paper if you could include both the reference link and a quote that gets the main point across for readers not in the mood to follow web links (or who are reading offline etc). The link to the preprint review is live, but the review is long and the pertinent part is not immediately obvious.

These links have been removed and text shortened to simplify the narrative (L59-64).

65- just to check, in this quoted text the reference numbers have been updated to match this paper's bibliography, right?

The reviewer is correct. This section of text has been thoroughly revised and can be found between L64 and 114.

85- I'm finding all this discussion of terms very confusing, since the block quote defines the terms and then you define them differently, but keep one of the WNWN definitions, but spend the next two paragraphs talking about how WNWN was sloppy with their definitions. It is hard to follow and feels pedantic and esoteric (and a little nit-picky of WNWN when everyone has been confusing these not just them). My suggestion to make this more straightforward: only include definitions in your own words, no quoted definitions. Then just say, this is the WNWN rarefying definition and what they called it, this is what I will call it, and this is my rarefaction definition. When people take notes reading these papers (and many new students just starting out with 16S analysis will), they will copy down those clear definitions, whether it's in the block quote or your text, and you want to make sure they are using the right ones. In turn they will make one note of "there's a lot of confusion in the literature of these terms." I think most people will not be interested in the long detailed listing of confusion, especially when it's just one paper being emphasized it feels like a personal vendetta instead of teaching. I know I've confused these terms before, and I think most people in the field will recognize the confusion as widespread with your few broad examples even if you make lines 85-90 and 95-100 more concise like the examples in 90-94. Also, preceding your new definitions (calling it subsampling), you might consider calling it "WNWN rarefying" each time it is discussed to clarify that you refer to that paper's definition.

This section of text has been thoroughly revised. I have removed the block quote and attempted to simplify the discussion.

117- I don't think it would actually be practical for differential abundance. Many of those algorithms are quite compute intensive. I know that for my most recent analysis, the differential abundance testing was the one analysis that required a server to run, so it's hard to imagine repeating that 1000 times.

The clause starting, “however, . . . ” at L129 indicates that the number of iterations and how to synthesize those results would need to be considered further. I appreciate that this is an important point to the reviewer, but it is beyond the immediate scope of my analysis.

129-typo grammar

This has been corrected

140- I got confused- how did 7 samples (line 131) turn into 26 samples?

WNWN used the fecal and marine samples to create OTU distributions and the other 19 samples in the GlobalPatterns dataset to create a distribution of sequencing depths. I have clarified the text at L139-145 and L156

258- wait how is 1.00 different than 1.00, 1.00, and 1.00?

This sentence has been corrected (see L276-281). The median accuracy values were 1.00, regardless of the sequencing depth. However, the confidence intervals varied widely.

330- do they give a reason for this choice? Ie do they say why they didn’t remove the 15% lowest OTUs from the other methods, too?

Unfortunately, they did not. The study from Weiss et al. (2017) also raised the point that this choice put subsampling at an obvious disadvantage compared to the other methods.

350- are they really skewed right “usually”? I have worked with very normally distributed read depths. This reference is just 12 studies, but even looking at them only some have long left tails, a few are more normal, and 2 even look skewed left to me. Also I’d note that this is a 3 year old pre-print that has not been published in that time, so perhaps not the most authoritative reference.

359- now there has been a leap from fig S1 of the old pre-print to a claim that library depths are log-normal. I don’t think there’s ample evidence (presented here) to support that, and I’m especially doubtful because it’s not my personal experience with raw sequencing data libraries

I would be open to considering other datasets, but in my experience sequencing depths are typically right skewed. It is important to note that the x-axis of Figure S1 from the preprint is on a log scale. While the distributions appear normally distributed if we don't consider the values on the x-axis, they are actually log-normally distributed. There are some datasets that have samples were fewer sequences, but these are relatively rare and typically point to a problem in library generation or sequencing. In reference to the fact that there's been a 3 year gap and perhaps things have changed. Again, this has not been my experience. For what it's worth, there's been a 3 year gap, in part, because the reviewers of that paper objected to my use of rarefaction in alpha and beta diversity analyses and general pandemic-related distractions. I continue to stand by everything said in that preprint

370- it might help to clarify how point #6 is different from point #8 because at first glance they're the same issue

Point 6 removes poorly sequenced samples whereas point 8 removes low abundance taxa. I am not sure how to improve the wording to make it more clear. I suspect confused readers will understand the difference when they read the body of the points.

376- I don't think it's the standard approach in dada2. The dada2 standard workflow only removes OTUs that occur as a singleton across the entire dataset (so 1 time in all samples). Deblur for sure does this though, and I'm not as familiar with Unoise. But I think dada2 does not, the dataset singleton cutoff is in my consideration pretty far from an arbitrary abundance cutoff.

I have removed reference to Dada2 and shortened the text since there is confusion over what it does under the parameters that most people use and what was originally published by Callahan et al. (see L398)

388- include citation again after "previous work" (I assume it's #4?)

A citation has been included to #4.

437- I mean, it is the traditional way to analyze 16S data, but don't you think it's a little outdated to only look at alpha diversity? And in WNWN they explicitly said, "Simulation B is a simple example of microbiome experiments in which the goal is to detect microbes that are differentially abundant between two pre-determined classes of samples." I know you are focusing on simulation A, but I would say this accurately reflects that alpha diversity as an end goal is somewhat outdated and

limited in its interpretation. This also goes back to my initial concern that rarefaction requires multiple recalculations of metrics, and that this could be really limiting to the downstream analysis options.

> In the Introduction, I refer to the traditional definition of rarefaction that was replaced by the confused definition that infected the field. The two references included in the Introduction (L46) and at L463 are from 1968 and 1971 and dealt with richness. It would seem that if one were to look at rarefaction (however defined) it should include richness given that rarefaction was originally applied 65 years ago to rarefaction. Perhaps it is outdated to *only* look at richness, but people still look at richness as part of their overall analysis.

497- this is a little misleading, because compositional approaches are used in downstream analyses to account for the intrinsic relative nature of sequencing data. They are not explicitly used to even out sequencing depth, that is done prior to the transformation. A log ratio approach, for example, could be applied to rarefied or normalized data. So they are not an alternative to rarefaction, they are a transformation (not a normalization) that enables a wider range of downstream analyses, specifically those that rely on Euclidean space.

527- there are so many methods that address compositionality when approaching differential abundance testing! You have kind of dismissed all of these, wrapping them up as normalization methods instead of data analysis methods.

> Some of the compositional methods certainly claim that they are invariant to differences in sampling effort (e.g., robust CLR). The data in reference 38 certainly demonstrate that adding psuedocounts prior to applying CLR methods supports the notion that the psuedocounts skew the distribution of the data. The text at L523-524 and L553-559

530- again, realistically probably not. These are time- and memory-consuming analyses and most people working with 16S data are not working on supercomputer clusters, or even servers for that matter.

> I understand the reviewer's concerns, but I disagree that this is too challenging. One need not do 100 or 1,000 iterations. Because most differences are likely to occur with rare taxa (L556-557), one could likely focus on the most abundant and significant taxa. Regardless, needing to use an HPC to analyze one's data should not be an impediment if it means getting the correct answer.

537- This needs the caveat that you're only looking at diversity. There are other methods (most obviously normalizing the data), and you have not addressed whether, when rarefaction is computationally not feasible, normalizing or subsampling would be better. Basically, WNWN compared those two, and found normalizing was better than subsampling. You have found rarefaction is better than both, for the limited case of diversity metrics.

> I have added a clause to the end of this sentence to indicate the the limitation of the current study (L564-565).

Reviewer #2 (Comments for the Author):

The author presents an investigation into the original conclusions of the influential WNWN paper (McMurdie & Holmes, 2014). Overall, I believe the author performed a service to the microbiome community by dissecting the WNWN conclusion that rarefaction is inadmissible. The manuscript is well written although tedious in parts and occasionally offering muddled advice. In general, it presents a case study in the extent to which (1) providing code is helpful for the progress of science; (2) science is challenging to get right; and (3) community correction plays an important role in the scientific progress. I commend the author for this effort. Below, I provide some suggestions for improvement.

MAJOR COMMENTS 1. I felt the focus on rarefaction contributing to error in microbiome studies misses the bigger picture about the effects of various error modes in microbiome research. There are numerous investigator choices that compound throughout typical microbiome experiments. See the recently published PMID 37076812 for several examples. I think this points at a bigger role for replication of samples through the entire process, and potentially with different protocols. I recommend adding discussion toward this end, even if this isn't the focus of the manuscript.

> I appreciate the reviewer's concern. I think it is widely appreciated that altering methods gives different results. I showed this here by the difference in read depth, clustering method, alpha diversity metric, beta diversity metric, shape of distribution. Given the length of the manuscript currently, I'm reluctant to add more text to the manuscript.

2. The figure legends are largely missing interpretation. Given the length of the manuscript, it would be helpful for the reader to have some way to easily understand what is the point of each figure. The colors are not consistent between the first two figures and their recreations in the

second pair of figures. Rarefaction should be clearly delineated each figure. Optionally, consider adding an initial schematic showing the different simulations, validations, and conclusions, as this was difficult to follow.

As suggested by the Editor, I have split the figures between the main body of the manuscript and the supplement. In selecting which figures to move to the supplement, I moved the first four "validation" figures. The first two of these used the color and symbol scheme used by the original authors. Since these figures are now in the Supplement and not immediately visible to the reader, I hope this reduces the confusion around the color choice. I have also revised the legend text to insure that they provide more interpretation. Finally, in the relevant figures, I have thickened the line to highlight the rarefaction data. Hopefully the thicker line and high contrast pink color will make these data easier to see.

3. I disagree with the author's conclusion in point #8 that very low frequency sequences should be considered in all analyses. There are, unfortunately, numerous sources of extremely low abundance reads, and the fact that they may appear in high depth samples is insufficient evidence of their authenticity. Some of these error modes can be eliminated by requiring a percent similarity to other (real microbiome) sequences, which would remove artifacts that aren't the correct amplicon. Perhaps suggesting a more nuanced approach that singletons (and their ilk) should be treated differentially depending on the analysis being performed, or some other conclusion intermediate to the position of treating singletons as real.

Requiring a percent similarity to known sequences will likely cause more problems than it solves. This is effectively the same as closed-reference clustering, which is challenged by the ability to detect novel diversity. Also, many singletons would have no problem passing a percent similarity threshold. The cited preprint (reference 4) deals with this topic in far more depth. My main point in this point is that the authors criticize arbitrarily removing sequences through subsampling, but think nothing of arbitrarily removing rare sequences without justification. Again, this is well beyond the scope of this study, but I continue to contend that it would be far better for people to generate high quality data instead of using single reads or pairs of reads that hardly overlap.

MINOR COMMENTS

1. In the introduction, please explain the meaning of the phrase “waste not, want not” in this context.

Thanks for this suggestion. I have added a brief description on the phrase to the Introduction at L53-55.

2. In my experience, 1k to 4k reads is still common when sequencing many samples on the MiSeq. This is contrary to one of the author’s numerous points.

The text describing point 2 has been revised to do a better job of indicating that smaller sized sequence collections are still observed (see L198, L278-279, and L280-282)

3. L332: “not known” to “not be known”.

This has been corrected.