I thank the editor and reviewers for their helpful comments on my manuscript. I have thoroughly addressed their comments. In addition, the revised manuscript has included alpha diversity estimates derived using the iNEXT R package, which combines non-parametric estimation and rarefaction to control for uneven effect sizes. I have also replaced the robust centered log ratio method, with the full Robust PCA method, which combines the robus CLR transformation with matrix completion. Regardless, the results are the same as robust CLR alone.

Reviewer #1 (Comments for the Author):

Within this manuscript Pat Schloss reports on the use of rarefaction to effectively control unequal observation efforts in 16S rRNA gene sequencing due to differences in sequencing depth. This paper is in response to a controversial idea within the microbiome community about whether rarefaction or other methods of controlling sequencing depth during diversity analysis are most appropriate. Overall, this paper shows convincing results that rarefaction is the most appropriate method for controlling unequal sampling depths at the present time. It was a pleasure to read and overall, I was happy with the manuscript. However, I do have a few comments that should be addressed by the author.

Major comments:

1. I found the result of increasing CoV in Bray-Curtis distances with increasing sequencing depth to be an interesting result given that increased sequencing depth should generally result in better confidence of what is within the underlying population. I'm wondering if the CoV results are difficult to understand due to the changing number of samples included at each sampling depth. Could this analysis be redone using only samples with sequencing depths that are above the highest rarefaction value chosen so that sample number does not vary across depth (this could be done in a subset of the samples/datasets presented). If changing sample number is not to blame are there any further speculations as to what is going on here?

> I have added a set of panels to Figure 6 indicating the standard deviation across the samples. This hopefully makes it clear that the standard deviation between samples is decreasing, but not as fast as the mean distance between samples. On the whole, the COV values are fairly small. I have added text to this effect at **LXXX**.

2. It is clear based on the results shown that treatment groups that are highly associated with variance in sequencing depth will result in high false positive rates during diversity analysis when

rarefaction is not used. However, I would argue that the simulation used does not in general reflect many natural datasets. It would be uncommon for most natural datasets to vary by treatment group by splitting sequencing depth via the median. I am wondering if this analysis could be done with varying levels of association between treatment group and sequencing depth. This will give the readers a better idea of how large this issue is in actual practice as the true association for most datasets may lay somewhere between figures 2 and 3. This would also help address how much of a concern this may be for previously published papers that do not use rarefaction during diversity analysis.

> I appreciated this suggestion and repeated the analysis with a less severe level of confounding. In addition to the perfect confounding of treatment groups by sequencing effort, I created an additional simulation where the smallest 5% of samples were assigned to one treatment group and the largest were assigned to another; the remaining samples were randomly assigned between the groups. The new results are in Figure 3 and the previous simulation are now in Figure S4. Although the results are less severe in the new simulation, they still affirm the effects of sampling effort on the propensity to get false positive results.

3. Based on the results of this paper it seems the key difference that is being demonstrated in these simulations as opposed to other works is the use of true rarefaction rather than subsampling. Further comments on this in the discussion would be useful to highlight why these results differ so significantly from other work such as the paper by Martino et al., in 2019 (PMID: 30801021) that introduced using robust CLR measures for microbiome data (and is quickly becoming one of the preferred choices of microbiome scientists). Furthermore, it would be nice to see whether the use of true rarefaction on the datasets presented in Martino et al., 2019 resolves the conflicting views between this report and Martino et al's findings on robust CLR measures. Overall, a clarification on how these two publications come to differing conclusions about best practices when comparing beta diversity would be helpful to a large audience within the microbiome field.

> I appreciate the reviewer's suggestion to look closer at the Martino et al study and their RPCA method. In the revised manuscript I used RPCA rather than the simple RCLR transformation. In the process of this investigation, I found a number of problems with the design and analysis of the benchmarking experiments performed in their study. While I make a vague mention of the differences between my analysis and the Martino et al. and McMurdie & Holmes studies in

the revised manuscript (**LXXX**), a more comprehensive analysis is underway. Unfortunately, it is beyond the scope of the current study.

4. I am wondering whether any of the differences in the results derived from studies such as Martino et al., 2019 (PMID: 30801021), can be attributed to sequence processing pipeline. While the author makes convincing arguments that removal of low abundance sequences may not always be appropriate, I am wondering how it would impact the results shown. Especially given the widespread use of this practice in the current literature as indicated by the author in the discussion. This would be especially important given the strong language around this practice in the discussion without primary data to back it up within this manuscript.

My responses to the previous points partially address this question. The data processing pipeline do not factor in for interpreting the results of this analysis since all of my benchmarking analyses used the same data for a dataset and each dataset was processed with the same pipeline. The methods don't know what type of processing was done upstream. Part of the value of using 12 different datasets was to provide diversity in community shapes. I could have used purely synthetic distributions as others have done (e.g. Martino et al., McMurdie & Holmes) and obtained similar results.

5. On line 164 the report indicates finding differences in richness between samples that should have the same OTU counts given the way the enrichment sampling was done (increasing sequencing counts not raw number of OTUS in the sample). It would make sense that metrics that consider abundances to change but not richness. Comments on this would be appreciated. Furthermore, this section jumps around between two different models of enrichment (abundance enrichment vs. raw count enrichment). A clearer separation of these two analyses would be appreciated to make parsing the text easier as in some cases it is unclear which enrichment strategy was used for the result that is being talked about.

The text in this paragraph has been edited to make it more clear what simulation is being used for each analysis. There is also text in the Discussion highlighting the value of using the Shannon or inverse Simpson's diversity indices over richness because they depend less on rare taxa (**LXXX** - "The primary alternative to rarefaction").

Minor comments:

1. Well, the current consensus is that sequence depth is mostly attributed to noise (Gloor et al., 2017. PMID: 29187837). It would be nice to comment on some recent papers that indicate that through certain methods of library preparation sequencing depth may be correlated with sample biomass and how this reflects with the current thoughts in this manuscript (Cruz et al., 2021. PMID: 33717032, Munich et al., 2022. PMID: 36396943).

> I have added the citations the reviewer suggested to the text at **LXXX** and fleshed out the discussion in this paragraph.

2. It would be nice if confidence intervals could be included in the correlation plots in Figure 1. Furthermore, if understood correctly Figures 2-3 y-axis should be the mean false positive rate over 100 dataset simulations? If so a measure of data spread would be appropriate for these values.

> The caption for Figure 1 indicates that the size of the confidence interval was smaller than the plotting symbol. As stated in the captions for Figures 2 and 3, “The false positive rate was the number of times a dataset yielded a significant result.” Therefore, it does not have a confidence interval.

3. While I agree that the rarefaction value chosen is not arbitrary and is usually a balance between keeping samples and having the highest sequencing depth possible (as indicated in the discussion of this manuscript). It would be nice to give stronger guidelines around choosing an appropriate value to highlight more robust ways of choosing rarefaction depths. One thought that came across my mind would be the use of CoV calculations from repeated subsampling (rarefaction) to determine appropriate sequencing depths. Or perhaps just a simple calculation of good’s coverage at various sequence depths (similar to rarefaction plots used now)?

> I’m reluctant to add more than what I already have at **LXXX** L281-289. In that text, I mention using Good’s coverage values as a check to make sure that the level of coverage is acceptable. In the revised manuscript I did add alpha diversity metrics that incorporate sampling depth and coverage to extrapolate (iNEXT) (**LXXX**). Although these methods perform better than their “raw” counterparts, there are still effects of sampling effort.

4. Based on these results and various studies on microbiome biases, alpha diversity metrics such as richness do not indicate the true richness of a sample due to a variety of issues included

sequencing depth. A comment on this in the discussion would be helpful, for those that are less seasoned in the microbiome field.

I have added a sentence to this effect in the Discussion at **LXXX**

Reviewer #2 (Comments for the Author):

This study addresses a controversy that has been playing out in the literature regarding the processing of 16S rRNA gene amplicon data. Some researchers contend that rarefaction (the random subsampling with replacement to account for differences in sequencing depth among samples) is undesireable because it “wastes data”. Schloss does a comprehensive analysis of how rarefaction performs compared to other ways of processing the samples (including using raw counts). He compares both alpha and beta diversity metrics using both a mock dataset and a collection of published datasets from a range of ecosystems. In nearly all cases, rarefaction

My only “major” critique relates to the core purpose of the article: to convince readers that rarefaction is a better way to process data than those proposed by other researchers. After reading the article multiple times, I still don’t yet comprehend why the other researchers came to their conclusions. At some level, they must have convinced themselves that the use of raw read count (or other methods) out-perform rarefaction. Did they really not do the same comparisons that are done here? Without reading those carefully, I can’t tell. Schloss does mention that they removed rare OTUs in their analysis. Is this the main reason why they did not detect the superior performance of rarefaction? I think just a brief explanation for Schloss’ interpretation of why his conclusions are different from McMurdle/Holmes would be even more convincing for readers who don’t want to dig too deeply into other papers. There is some discussion in the introduction, but I’d like to see it in the actual discussion section, after the full presentation of the new analysis.

Kudos. Love the running title. Tell us how you really feel!

My pleasure :). I have added text to the closing paragraph of the Discussion (**LXXX**) to elaborate on why the results of my benchmarking studies differ so strikingly with those that have preceded mine.

Minor comments

Line 69 - This implies that McMurdle/Holmes used an incorrect interpretation of differences between rarefying and rarefaction. This is an opportunity to clarify! Do you agree with their

interpretation (but want to promote the use of rarefaction specifically)? Line 72 - Do you mean that the reanalysis did not remove 15% of the samples? Line 72 - Change "method" to "methods"

I have clarified this paragraph of the Introduction (**LXXX**). My paper describing the re-analysis of WNWN has recently been accepted for publication at *mSphere* and goes into far greater details regarding the problems with the original study.

Line 78 - Did they use the term "toy" in their work? This comes off as unnecessarily snarky. Would it be appropriate to use "simulated" instead?

No snark was intended. "Toy" is jargon in the field for a simplistic example desined to make a point. To avoid confusion I have used "overly simplistic" instead.

Line 94 - Delete one of the "from"s

This has been edited

Line 113 - I would say that the Raw and SRS normalized are completely insensitive to seq depth when using inverse Simpson, based on Fig 1A? Not sure this matters in the end, but be transparent in the explanation of the results.

This text has been clarified

Line 167 - Were the 3% of OTUs selected at random, or did you remove the rarest 3%?

This text has been clarified

Line 194 - Change "increased" to "increase"

This text has been corrected

Line 210 - Drop your excessive sig figs . . . . Use "101" fold

The number of significant figures has been corrected

Line 281 - I think this should be "my personal process for selecting the rarefaction threshold"?

This text has been corrected