I am grateful that the reviewer sees value in this work and the questions I have raised. I apologize for taking so long to resubmit a revised version of the manuscript. In part because of the reviewers' comments, I paused work on this project to take on the question of whether the use of rarefaction was valid for alpha and beta diversity metrics. Two manuscripts: (1) a direct response to the original McMurdie and Holmes study and (2) an approach using a different simulation framework have now been posted as preprints to bioRxiv and submitted for review also at *mSphere*. Although the Editor's comments indicated that it was not necessary to take on rarefaction, I felt it would overcome many of the concerns that the reviewer and others have raised. Beyond the issue of rarefaction, I hope that the reviewer will find that I have addressed their comments throughout the manuscript and taken to revise the text to make my methods more clear and results more convincing. The line numbers included below correspond to the "track changes" version of the manuscript.

A significant challenge of this type of work is that we do not know the *true* answer for any of these communities. Any analysis that has been done regarding thresholds for removing sequences has been based on mock communities which are highly artificial. Aside from analyses based on mock community data there has been little to no rigorous work demonstrating that we *should* be removing singletons. That decision has made by fiat rather than based on data. For example, McMurdie and Holmes decrie removing samples and sequence data by rarefaction, but thought nothing of removing rare OTUs in their analysis. Thus, I feel that I am left arguing against a position that really has no published research supporting its use that approaches the level of analysis done in this study.

I read the manuscript "Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data" from Dr. Schloss with great interest. There are a diversity of opinions about whether ASV or OTU tables should or should not be filtered or quality controlled prior to subsequent statistical investigation and an even greater diversity about the specific approaches that should be applied. This may hold important implications for whether and which patterns are observed in the analysis of microbiome data, yet there has been surprisingly little work to measure what the effects of these decisions actually are. Dr. Schloss' effort to evaluate the effect of these decisions in a data-driven way thus holds broad implications for the entire field of microbiome data analysis. In particular, Dr. Schloss analyzed a variety of publicly available datasets to determine how the removal of rare ASVs or OTUs affects the patterns present in microbiome data. This extensive analysis seeks to clarify whether the common approach of removing rare variants - which are often presumed to represent spurious sequences (i.e., artifacts) - from an analysis is meritorious. Dr. Schloss' analysis points towards two big conclusions: (1) rare variants add information to microbiome studies and should be retained and (2) rarefaction techniques should be used to control sequencing depth and thus the number of artifacts present in an analysis.

While these are important questions to address, I do not walk away from the manuscript - in its current form - convinced by its conclusions, especially given some of the strong language that is invoked. In particular, I think the manuscript could make a better case by improving its clarity in some areas, especially with respect to methodology and rationale, and by revising or adding analyses in other places to solidify conclusions. Moreover, some of the conclusions may require hedging given caveats that are not currently considered in the present analysis. Given that this manuscript upends long-standing practice and expectations, it seems especially important that it fill in these gaps.

Below, I itemize my criticisms with the manuscript. I hope Dr. Schloss sees these as constructive comments designed to help him improve this manuscript, which I ultimately think will play an important role in the advancement of microbiome data analytics.

1. It is currently challenging to evaluate the work because the methodological details used to conduct it are frequently sparsely described. The materials and methods are in effect integrated into the results section. While I do not have a problem with this per se, it seems to have come at the cost of not providing specific details about various analytical steps that were implemented as part of the analysis as well as the rationale for implementing these techniques. This ranges from relatively straightforward aspects of methods, such as the reason for why these specific data sets

were selected (from among the others that fit the same criterion of targeting the V4 locus with the same sequencing technology), to whether multiple tests correction was applied to statistical tests. In other places, the methodology is confusing as described, especially regarding the various null model sampling approaches (for example, see my specific comment about line 200 below). While I appreciate that Dr. Schloss has made all code available to readers, I think it's important to ensure that an understanding of the approach by a reader does not depend upon their ability to read code. I therefore recommend that the technical approach be more explicitly described in the manuscript.

I appreciate the reviewer's concern about the mixing of methods throughout the Results section. I would prefer to leave much of the simulation descriptions with the results to make it easier for the reader to understand what is going on. Clearly, I needed to do a better job of describing those connections and I appreciate the reviewer for pointing out places that could be improved. * For the various simulations and analyses, I have tried to introduce the analysis by describing the methods more clearly. With this revision and thanks to the suggestions from the reviewer below, I hope that the descriptions of the simulations are more clear. * I have also tried to improve the description of the statistical tests and thresholds I used with the simulations. All of the comparisons were binary within each study and did not require use of corrections for multiple comparisons. * There are certainly more datasets that could fit my criteria, but I felt that 12 diverse datasets were sufficient. I would be happy to include additional datasets that the Editor or reviewer recommend. Ultimately, the actual datasets that are used is not critical as I am only using the datasets to simulate various community structures. I feel that the description of why datasets were selected is clear in the "Datasets" subsection of the Results section and the "Data curation and analysis" subsection of the Materials and Methods section.

2. I'm not sure that the manuscript provides evidence that rare sequences should be retained in an investigation, but rather documents the sensitivity of commonly applied analytical approaches to the decisions made about whether and how rare sequences should be incorporated. The former statement requires demonstrating that the inclusion of rare variants improves analytical accuracy. While the present analysis suggests that including rare variants impacts alpha- and beta-diversity metrics as well as the sensitivity of tests that utilize these metrics (according to the null model approach), it is not evident if the filtered or unfiltered data better reflects the underlying ecology. As a brief aside, Figure 5 offers the clearest insight into the potential impact of filtering on analytical accuracy, but it is not clear from the methodology how treatment groups were established and thus it is hard to follow Dr. Schloss' interpretation here.

As mentioned above, a consistent challenge in studies such as this one is our lack of knowledge of the true structure of microbial communities. Although Mock community data can tell us what is spurious and what is real, their community structure is highly artificial and unlike anything we see in nature. Therefore, we are left to use model communities where we know their true structure and perform simulations based on those models. Of course, "all models are wrong, but some are useful" (George Box) and we are left to ask whether the models I've created here illuminate concerns that we should have with removing rare sequences. The models presented here do not incorporate sequencing errors per se, rather rare ASVs that may be good or spurious. What we find (and the reviewer notes) in Figures 4 and 5 is that removing rare sequences causes problems for statistical power (Figure 4) and false detection of differences (Figure 5B). My models argue that retaining rare ASVs provides a better sense of the ecology of the study because it makes it easier to detect real differences (i.e., power) and harder to detect false differences.

In general, much of the interpretation of the results appears predicated on the idea that the rare variants represent real taxa. However, there is only limited evidence in support of this view. In the section "The nature of singletons", Dr. Schloss finds that singletons are more prevalent in more shallowly sequenced samples and that singletons in a sample are often also present in another sample. Dr. Schloss concludes that these singletons are thus likely real variants and the rest of the investigation seems somewhat predicated upon this interpretation. However, I am left with several questions:

(A) Are the correlations meaningfully representative of a trend across all samples, or are they driven by a small number of outlier samples? Providing support of these trends by plotting the scatterplots

for each study would help assuage this concern.

I appreciate the reviewer's concern that the correlations presented in Figure 1AC could be driven by outliers. It is worth noting that I used Spearman correlation coefficients, which are robust to the effects of outliers since it is a ranked-based test. As suggested, by the reviewer, I created two new figures, Figures S2 and S3, which shows the requested plots for the data in Fig 1AC. There are no obvious outliers in these figures that are skewing the Spearman correlation values.

(B) Despite these observations, couldn't these rare taxa include artifacts nonetheless and, if so, do real rare variants and artifacts contribute uniform information in all subsequent analyses? An effort to directly measure the impact of spurious sequences on analytical accuracy would help offset this concern.

Rare taxa could certainly be artifacts. At L264-267 I discuss how even under the best of circumstances PCR artifacts like chimeras and sequencing errors will persist. As part of this discussion I described two strategies for dealing with lingering artifacts: rarefaction (L268-272) and making comparisons between treatment based on relative differences in diversity rather than absolute quantities (L273-283).

(C) Since the processes underlying the generation of artifacts are not explicitly described, the reader is left unsure about the expectation that the formation of singleton artificats is driven by sequencing depth.

I hope that the revised text addresses some of these concerns.

3. I'm not sure that the description of variant filtering accurately reflects common practice. For example, while many groups appear to apply count thresholds to determine which taxa should be retained for analysis, these thresholds are typically applied across all samples and are often cross referenced with prevalence filtering. As a result, taxa that are singletons in one sample but present in others are typically included in an analysis despite otherwise being below the count threshold. In Dr. Schloss' study, however, such taxa are considered candidates for filtration, which raises questions about the applicability of the results. Moreover, frequent practice thresholds on relative abundance rather than direct counts to offset the potential for sequencing depths to drive singleton formation effects. Dr. Schloss may consider discussing or even experimentally measuring the effect of these alternative strategies on the outcomes of his analyses.

- As outlined in the Introduction, sample-by-sample abundance filtering is the norm for dada2, Deblur, and the Bokulich pipeline whereas study-wide abundance filtering is used by UNOISE2.
- [simulation?]
- study-wide abundance filtering
- counting the number of samples that have a sample
- The reviewer suggests using relative abundance thresholds as an alternative to abundance thresholds. Relative abundance thresholds are effectively the same as rarefied OTU counts since, mathematically, relative abundances are the average number of counts for the OTU across a large number of resamplings divided by the total number of sequences in a sample. Unfortunately, filtering by relative abundances do not control for the total number of sequences in a dataset and so any ASV below the threshold will effectively have its abundances redistributed across the other ASVs. This is the same problem I encountered when removing rare ASVs. [simulation?]

4. Related to the above, it is not clear if the results presented here are generalizable to other investigations. For example, it is possible that the use of the high-quality data sets included in this analysis or the specific bioinformatic techniques applied here results in a lower than typical rate of artifact formation compared to what many readers may encounter in their own analyses. For example, studies that do not apply library prep approaches that produce two overlapping V4 reads, but that rather amplify a longer amplicon target and thus have lower sequence accuracy, may have an increased rate of forming artifacts in their data. Likewise, ASVs in this study are sequences that tolerate a 2 bp difference, which may differ from the set of ASVs produced using another approach.

Since the study never directly measures the cost of including artifactual sequences (2B above), it is not clear if the recommendations made here are meaningful for investigations that involve data subject to a greater rate of artifact formation.

I think the reviewer may be putting too much emphasis on the importance of the individual datasets. These datasets were selected to help me simulate communities with different properties, number sequences per sample, and number of samples per study. At the same time, I would resist any attempt to normalize the generation of data that we know has a high error rate. Why shouldn't researchers be expected to generate the best data they can?

5. Conclusions around rarefaction need more justification or support. There are several strongly worded statements in the manuscript about the use of rarefaction techniques, but there appears to be little direct assessment of the impact of rarefaction on study outcomes. Additionally, there is no comparison with alternative rarefaction-free normalization techniques, including simply filtering based on relative abundance. Given the on-going debate in the literature about the value of rarefaction, it seems important that the manuscript solidify this point if the language is to be included as is.

As I mentioned in the preamble to my response to the reviewer's comments, since receiving their comments, I have written two manuscripts using different simulation approaches to demonstrate that rarefaction is critical for drawing proper inferences from microbiome studies. As outlined in the response to McMurdie and Holmes's study, their approach was effectively designed to make look rarefaction look bad - this was principally achieved through (1) removing 15% of the samples and then penalizing the method for the loss of those samples and (2) basing the analysis on a cluster analysis of samples that used a clustering algorithm that performed the worst of the available options. In the second manuscript, I developed a different simulation strategy that is similar to the one used in this study and employed additional normalization approaches. Both studies show that rarefaction is critical for controlling for uneven sampling effort when pursuing alpha and beta diversity analyses. I contend that the results I have shown regarding coefficients of variation and type 1 and 2 errors support the use of rarefaction. Finally, the reviewer encourages the use of relative abundance, which I show in my rarefaction papers can lead to problematic interpretations.

6. The rationale underlying some of the analyses, especially with respect to the null models, needs support to help readers connect the results with the conclusions. For example, for the section on line 122 ("The impact of removing rare ASVs on the information represented in each sample.") it seems self-evident that removing rare taxa will affect the estimate of alpha-diversity, especially when considering the well documented rank abundance distribution and long-tail of low abundance taxa that typically comprises microbial communities. Dr. Schloss may get more mileage out of this section by clarifying why this analysis is noteworthy, or by focusing of metrics that are less directly sensitive to the number of taxa (perhaps evenness).

- Sobs is definitely obvious
- Bigger point in Figure 3 is that the coefficient of variation between samples goes up with removal of rare sequences reducing power to detect differences in communities
- Will also calculate Shannon evenness (keep in mind that this is dependent on richness)

Additionally, I'm not sure that the treatment effect simulation analysis is as meaningful as indicated by the text (though again, I'm not sure I follow the approach properly). If the ASV sampling procedure to determine which ASVs are increased in abundance is simply selecting from a list of ASVs, then the long-tail of rare ASVs will tend to be selected versus those that tend to abundant. Consequently, when the procedure amplifies the abundance of these (and only these) selected ASVs, again which are biased towards rare taxa, then it will be rare taxa driving the differences between simulated treatments. Abundance-based filtering of said taxa will obviously have an effect on the analysis. Is this akin to what happens in microbial communities in nature? Well, maybe in some instances, but lots of studies find that it's the abundant taxa that differentiate treatments. Such studies may not need the singleton ASVs to observe the effect. Could those ASVs help clarify the effect? Perhaps. But inclusion of artifact sequences (which again are not directly tested here) could theoretically increase the error of the analysis in such cases at a greater rate. I would recom-

mend considering a parallel analysis that simulates ASVs based on their abundance to help clarify whether the interpretation drawn from this section is relevant to communities that behave along the lines described above. Alternatively, hedge the conclusion by clarifying the above caveat, but that this analysis may yet reveal insights into how rare ASVs could reveal differences currently being missed in communities that do not substantially change with respect to the abundant fraction of community membership.

- As the reviewer points out, we don't know. There will certainly be times where differences between communities are driven by rare populations and times where it splits by abundant populations. We really don't know. Those that are driven by abundant populations will be obviously different already so it's a moot point for that case
- High quality data reduces risk of artifacts and rarefaction allows us to control for inclusion of artifacts

Minor comments:

Line 94: What is meant by "relatively small" when referring to the marine and soil data sets? Based on table 1, it seems to be the number of samples. But you can help readers by being more explicit here especially given that the surrounding context of discussion pertains to sequencing depth.

The reviewer is correct that I was referring to the number of samples. I have clarified this sentence.

Line 102: "In addition, with the exception of the samples from the marine and sediment datasets (Spearman correlation, P>0.05)..." Indicate parenthetically here that these samples were non-significant (as opposed to the immediate assumption that they are positively associated) and move the parenthetical pointer to spearman and the pvalue to later in the sentence.

The sentence has been edited as suggested by the reviewer.

Line 148: "For example, after this procedure, each of the 490 samples from the human dataset would be expected to have the same richness and diversity of ASVs and one would not expect to find treatment-based effects between the samples." It would be more compelling to demonstrate that this expectation was met in the resulting distributions. Can you offer graphical or statistical evidence in support of this?

This was shown in Figure 3 when rare sequences were not removed (i.e. when the smallest number of sequences per ASV was equal to 1). In this case the coefficients of variation were nearly zero. Furthermore, in Figure 5A and 5B the Type 1 errors were near the expected 5% (i.e., P=0.05) for the case when the smallest number of sequences was equal to 1.

Line 175: How many samples are being selected for each simulated treatment in this analysis? At first I thought all samples in a study were selected, but it looks like this isn't the case, as an undescribed number of "communities were randomly sampled from the null distribution..."

The simulations did indeed use all of the samples. I have clarified the text to indicate that all of the samples were randomly assigned into one of two treatment groups

Line 179: What p-value threshold is being applied to "detect a difference" and quantify the "fraction of significant tests"?

I have clarified this sentence to indicate that a significant test was one where the P-value threshold was 0.05.

Line 200: It's a little challenging to determine what happened here: "First, I sampled sequences from the null distribution for each dataset and randomly assigned each sample to one of two treatment groups and determined the richness and diversity of ASVs and OTUs". For example, what is meant by "sequences" here? ASVs? How did you sample the sequences? Are you saying that you subsampled the data somehow? I recommend you clarify.

I have restructured the text here to make it more clear how the simulations were created.

Line 204: What is meant by "meaningfully vary"? Was this quantified in some manner?

I have inserted the median, minimum, and maximum p-values that were observed across the range of conditions that are listed in these sentences.

Line 215: "However, when rare taxa of any frequency were removed, the probability of falsely detecing a difference as signifiant increased with the definition of rarity". But Fig 5B seems to show that the rate only tends to increase when singletons are removed, such that the rate is flat as more abundant rare sequences are removed.

The reviewer is correct that the curve is flat; however, the curve flattens at a very high level when any rare sequences were removed. That's the problem. The value is the lowest when all sequences were retained. The reviewer has correctly identified the problem: if any sequences are removed, the false detection rate increases dramatically.

Line 236: "The practice of removing rare sequences from samples seems to be a response to researchers prioritizing the number of reads and length of sequences over their quality." I'm not sure this is entirely accurate. Some of it stems from this, but it also appears to have emerged as a way to control for overclustering, which often results in spurious OTUs. This methodology may have been a holdover in the era of ASV analysis, where it may not be as critical.

This sentence has been modified to reflect that the perception of overclustering is a product of prioritizing the number of read depth and length of sequences. Again, we don't know what is too many OTUs if we don't know the true number of OTUs in natural samples

Line 250: "In practice, I suspect this range is actually larger since researchers may have opted against depositing samples with fewer reads into the SRA". Perhaps. But this would imply that such samples were not (nor would be) included in the analysis and not subject to the bias being discussed here.

This sentence has been removed in the revised manuscript

Line 255: I think I follow your intention, but note that your results of finding that rare sequences are higher in shallow depth samples seems initially at odds with the subsequent part of the discussion that observes that "number of spurious sequences increases with sequencing depth". You can avoid some confusion by using this observation is the expectation and then discussing how your results differ from this expectation, reinforcing the interpretation that the rare sequences observed in your analysis are real variants. This may be your intention in the current narrative; I simply recommend a modest restructure of this logic to improve comprehension.

As suggested by the reviewer, this text has been revised to make it more clear.

Line 287: Typo. Change insure -> ensure

This has been corrected in the revised manuscript