Reviewer #1 (Comments for the Author):

The contribution by Sze and Schloss describes the bias associated with PCR-based amplicon sequence analysis and how it varies due to the use of different polymerases and PCR cycle numbers. The analysis is based on V4 primers, Illumina MiSeq, and processed with mothur.

1. The manuscript is well written, with the goals well described. Certainly the contribution makes a compelling case that, in mock communities, there are measurable differences in the amount of chimeric sequences and error rate across different polymerases (in line with manufacturers specifications for errors). However it is less clear to what extent these measurable differences are ecologically or biologically relevant. In fact, the manuscript presents some evidence to the contrary and indeed the authors waver back and forth on this important point throughout the manuscript. Clearly, the "truth" lies somewhere between the argument that samples amplified with different polymerases and number of cycle conditions are impermissible to compare (as line 27 seems to declare) or that the difference can be ignored altogether (clearly not the case). It is likely that certain applications, e.g., comparing the alpha diversity, could be explicitly discouraged, given the evidence reported, whereas beta-diversity patterns are relatively more robust. In my view, the results here should be used to help understand potential differences between studies and within meta-analyses, explain their limitations, and constrain their interpretation, rather than disallow them.

> We have added text and two citations at L280-283 of the marked up version of the manuscript to clarify our position. We are concerned that one outcome of this type of research is the idea that people could pool samples from across studies to make comparisons. The interactions are too complex to ever get an adequate understanding. Instead, researchers need to compare effect sizes from similar studies to determine whether results replicate or they need to control for experimental differences by treating the study as a random variable in a mixed effects model.

2. I think it would be beneficial to have a table that explicitly reports the trade-offs between the different polymerases, empirical error rates, cost per reaction (if possible), processsitivies (see below), and chimera tendencies shown. Perhaps one will come out way ahead, or perhaps trade-offs will be apparent.

> Given that we already have 6 figures we are reluctant to add another figure or table to the body fo the manuscript. We have expanded our discussion of polymerase choice at L289-295 of the marked up version of the manuscript. Also, the goal of the paper wasn't to make specific

1

recommendations regarding polymerases, but to define a set of parameters to consider when selecting a polymerase.

Technical details:

3. Good practice would include loading equimolar amounts DNA templates to each PCR reaction. Best practice would probably be quantify the template via a qPCR on the variable region of study, before the amplicon sequencing reaction, but Pico Green or Qubit is commonly used. It is unclear if this is routinely done via the Wet Lab SOP. If not, there are likely to differences in how, e.g., dNTPS are exhausted throughout the PCR, ratios of metal co-factors (Mg) to template, and ratio of primers to template. All these factors could have implications to the number of chimeric sequences, for example. While this shouldn't be an issue for the mock communities, and probably not for the natural samples measured here (which were easily distinguished), it is very relevant to the topics at hand (chimeras, error rates, cycle numbers, alpha diversity, etc) and seems it should be included in the discussion. Please note the following reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5853488/.

> We have added a reference to the paper by Multinu et al. (2018) and a comment about balancing the template concentrations across samples at L305 and L312 of the track changes version of the manuscript.

4. Polymerases have different processivities (the number of bases added before disassociating) or may even come in different working concentrations, or units/mL, or some polymerases may just achieve higher amplification efficiencies in given conditions (i.e., combat inhibitors). This is an underlying factor that may complicate comparing the different polymerases at the same number of cycles. It may not be the number of cycles of PCR that is performed is the most important, but rather the number of amplifications per template. It also may explain the number of chimeras observed. I think a discussion is warranted. One would be able to determine how efficient the polymerases were, in practice, by inspecting the agarose gels that were run on 12-24 samples per 96-well plate (according to the the methods). Otherwise, this information will be lost because normalization plates were used on the post-PCR reactions, thus the post-PCR quantification will never be known, and how the processivity varied between the different polymerases in this experiment would be lost.

> We appreciate the reviewer encouraging us to take a more mechanistic look at the results. Although this information is interesting, we feel that it goes beyond the scope of the current

manuscript. Ultimately, researchers pick a polymerase and use it according to the manufacturers suggestions. This is not always ideal, but it is the reality. We have not optimized the reaction mixture for every polymerase and have instead chosen to analyze the polymerases as sold by the manufacturer using their suggested protocols.

5. Please state the E. coli positions of the V4 primers, as well as the primer sequences, in the methods. (It takes at least 5 clicks to find this very important information, and the paper referenced in the SOP seems to list the primers in a rather complicated Figure 1, and a paper that tests a lot of primer combinations which is confusing).

We have added a sentence at L356-358 of the track changes version of the manuscript to clarify the primer sequences and the locations where they anneal.

6. It would be highly informative and more broadly applicable if the chimeras are described. Here are some questions I have with various degrees of applicability to this study (not all need to be addressed, but some merit discussion in the paper). - What chimeras that are formed? - Do the chimeras form at re-producible locations? - Are they random? - Are there many tripartite chimeras "trimeras?" More? - Could these observations have implications to other popular amplicon regions? - Are there chimeras formed between distantly related taxa in mock communities or is it just between closely related taxa? What are the implications of this? - Related, would the Zymo mock community be expected to over estimate or underestimate chimeras, based on the number of close relatives compared to a natural community?

We have added a brief discussion at L140-151 to outline the characteristics of the chimeras. We found that a significant number of chimeras were replicated across all of the polymerases and that within each polymerase, these chimeric sequences represented a majority of the chimeras. From our data, it appears that the formation of chimeras is driven more by the characteristics of the sequences than those of the polymerases. Because we didn't test other mock communities, we are reluctant to make broader claims about the relationship between sequence and community composition with chimer formation. Given chimeras are more likely to form between related sequences, it is likely that communities with low genetic diversity would form more mock communities.

Other details:

7. Figure 1, add more resolved y-axis for the denoised contigs rate.

We have made the second panel of Figure 1 have its own y-axis scaling to improve the vertical resolution.

8. It would help if, in the figures, it was somehow more clear which data corresponds to mock communities and which are for natural samples.

We have added text to the figure legends to clarify when sequence data were generated from the mock community.

9. Figure 5, don't use red and green together due to common red/green colorblindness.

We have adjusted the color palettes in Figures 4 and 5 to be more red/green colorblind friendly.

10. Figure 5, is there statistical support for the circles? If not, they are misleading, and seem arbitrarily drawn (also for Figure X)

The ellipses represent bivariate normally distributed confidence intervals. We have added text to the legends to make this clear.

11. Should be "KAPA" throughout not "KAPPA", no?

The reviewer is correct and we have corrected the spelling throughout.

12. Line 286: re: ASV pipelines, "Although these methods. . . risk . . . merging 16S rRNA gene sequences from different taxa into the same ASV." This is also, of course, the case with OTUs, not sure why this is relevant, should be less of an issue, in theory, with ASVs.

We have removed this sentence from the Discussion.

Reviewer #2 (Comments for the Author):

Sze and Schloss compared the performance of high fidelity polymerases and varying number of rounds of PCR using the LGC bacterial community standard (with replicates) and 4 human stool samples. The authors found that perfoming more PCR cycles increased deviation from the expected composition of the mock community. Variation in human stool samples was larger than that introduced by the choice of polymerase or number of PCR cycles. Some findings described in the ms may be of practical interest for other laboratories, such as choosing the DNA polymerase

with lower error and or chimera rates. I have a couple of (rather minor) concerns that the author may wish to consider:

13. Line 49. Could the authors mention emulsion PCR (PMID 16791213) in the context of reducing chimera formation?

We have added a citation to the paper by Williams et al. (2006).

14. Line 113. Could the authors describe how exactly they calculated the error rate?

This is outlined at L374-376 of the track changes version of the manuscript.

15. Line 139. Please describe how specificity/sensitivity were calculated.

This is outlined at L380-383 of the track changes version of the manuscript.

16. It may be interesting to indicate the range of Bray-Curtis distance between different polymerases for the same number of PCR rounds.

The range of distances between polymerases across the number of cycles has been included at L223-224.

17. Line 258 "Among these polymerases, the Kappa polymerase resulted in the lowest error rate, lowest chimera rate, and least bias across rounds of PCR." This is not evident from Figures 1 and 2a, and is valid only for 35 cycles.

We have clarified the language to indicate that across the cycle conditions we considered for both the mock and human community data KAPA performed the best and that these benefits were most accentuated at 35 cycles (L291).

18. Line 259. correct "the the".

This has been corrected.

19. Line 173. Q5 was stable across rounds of PCR and this can be stated. Also indicate that Kappa was the only polymerase with the ratio >6 for all number of rounds (not only for 25-35, but also for 20 cycles).

The text has been updated to better state both observations.

20. The data related to 20 rounds of PCR are lacking in many figures. For example, Accuprime polymerase/20 rounds is given in Figure 2 but absent in Figures 2A and 2C. Also in 5A and 5C some 20 round PCR data are missing.

The 20 round Accuprime data has been added to Figure 4. We appreciate the reviewer's careful eye in detecting this omission. We have indicated in Figure 5 that it was not possible to obtain sufficient data for all human stool samples for Accuprime and KAPA at 20 cycles.

21. Legend to Figure 1. Please indicate that these data refer to the mock community (check other figure legends).

We have clarified the use of the mock community data vs. human stool samples throughout the figure legends.

22. Figure S1. The label on y-axis is not easy to understand. This could also be clarified in the legend. Please explain what the grey line corresponds to. Did the authors always represent mean value of 4 replicates? If so, indicate this in the legend (also in other figures).

This has been clarified as the reviewer suggested in the next comment. We have clarified where the mean of 4 replicates were presented in the figures.

23. Line 551: For clarity, please replace "The two S. enterica V4 sequences' by, for example, 'Dominant and rare variants of the S. enterica 16S rRNA gene sequences'

This has been clarified as suggested.

24. Line 546, correct 'the the'

This has been corrected.

25. Line 341. Indicate 'rRNA operon'

We have clarified that we are describing the *rrn* operon.

26. In Figure 5C, symbol description: for different number of PCR cycles use lines of different colours, instead of circles (Accuprime is represented by circles).

We have replaced the colored circles with larger square swatches for each color.

27. Figure 4A, symbol description: "errors and perfect chimera removal" and "errors and chimera

removal with VSEARCH" is slightly confusing to me. Did the authors mean "no error removal"

> The legend for Figure 4A and the associated text (see L205, L212, L215) has been edited to
> clarify what was done and how each treatment relates to the description in the manuscript.

28. The text in the Results section Lines 180-191, Figures 2 & 4 and "Sequence processing" in the Methods section should be modified and better related to each other for easier understanding. For example, Figure 4A shows three different ways of chimera/error removal but this cannot be easily related to the text in Methods, Results and Figure legends. A new figure or table may be provided, that briefly describes (even schematically) the three methods (possibly with an acronym for each of them that will be used in the main text and figures) including the tools used. The reader would better understand when and why UCHIME and seq.error were used, and what 'our method', 'perfect chimera', 'true chimera', etc. correspond to.

> Please see our response to the previous comment from Reviewer 2.

29. Discussion. The authors found that effects of PCR cycles were smaller than inter-individual stool differences. Earlier study (PMID22300522) showed smaller effects of PCR cycles than those introduced by the choice of the hypervariable region sequenced, which the authors may mention/discuss in the ms.

> We suspect that the earlier result found differences by variable region because the sequence
> reads did not fully overlap for the longer regions. As we have previously shown in Kozich et
> al. (2013), this would exacerbate the error rate, which would be compounded with the number of
> cycles.