Reviewer #1 (Comments for the Author):

The manuscript by Patrick Schloss et al provides an updated on the status of the microbial census. By microbial census the authors refer to the census of bacteria and archaea assessed through the full length 16S rRNA gene sequences.

The manuscript is well written and the authors have done a good job in summarizing the current status of the census and the bottlenecks in scaling up.

We appreciate the reviewer's kind words regarding our manuscript as well as their suggestions that we have done our best to incorporate into the new version of the manuscript.

A few minor issues to consider below Page 1. While the manuscript describes a census of bacteria and archaea only, it is referred to as microbial census which however includes microbial eukaryotes. As such the word microbial in the context of this paper and this census is misleading. I assume that the authors would prefer not to use the word prokaryotic census, but microbial is not accurate either.

Where appropriate, we have made the text more specific to describing the "archaeal and bacterial census" rather than the microbial census.

Lines 22-24: The meaning of the sentence "Surprisingly, we have done a relatively poor job of keeping track of the ongoing effort to characterize the biodiversity as represented in full-length 16S rRNA genes", is a bit confusing. There are several databases that keep good track of all the available 16S sequences so the authors should clarify here what they mean.

We have edited this sentence to hopefully make it more clear that we're referring to the fact that the ongoing collection and deposition of sequences in public databases is somewhat haphazard. In contrast, our analysis is more formal and allows us to take stock of the quality of the biodiversity within those databases.

Lines 25-27: The authors claim that "We found that the ongoing effort has done an excellent job of sampling the most abundant organisms, but struggles to sample the more rare organisms". However, recent literature is not very much supportive of that. In a recent publication (Eloe-Fadrosh et al 2016) it is shown that there is a large number of unaffiliated 16S sequences by looking at the assembled metagenomics sequences. Since typically, most of the 16S sequences that get assembled from metagenomic datasets are not of very low abundance organisms, perhaps the statement above doesn't seem to hold up? Furthermore, the authors agree in the discussion that the primers uses for PCR have biases, which contradicts the statement that ongoing efforts have done an excellent job in sampling the most abundant organisms.

This observation fits with the overall story of our analysis. In the Importance section, we are referring to the large number of singleton OTUs in the database. It is not hard to conceive the problem described by Eloe-Fardrosh when one combines the limited environmental sampling that has been performed.

Lines 142-143: the sentence "Based on the representation of sequences within the SILVA database, in 2006 there were 61 bacterial and 18 phyla" seems to be incomplete. Do they mean 18 "archaeal" phyla?

This has been corrected.

Lines 253-256: The authors discuss the contribution of assembled metagenomics data in showing the presence of introns in 16S rRNA genes. They don't discuss at all the 16S primer mismatches though as another source of significant diversity missed even from abundant organisms as has been recently shown.

This is actually described in the preceding sentence.

The authors are also using throughout the text the term "zoonotic" as one of the broad environmental categories. The use of this term is confusing. Do the authors mean host-associated? If yes, I suggest to adopt this term, since this is the term used in the SILVA database and the term recommended from the Genomics Standards Consortium and the MIxS standards. Zoonotic also refers to pathogenic/disease states and unless the authors refer specifically to this type of environments rather than host-associated environments, I would recommend changing the term here.

We have changed "zoonotic" to "host-associated."

Also, while Table 1 refers to Zoological samples, one of the curves in Figure 1C refers to Zoonotic. I assume that the zoonotic in the figure corresponds to the zoological in the Table, but keeping two terms is confusing, I suggest the authors will also use zoological in the figure to match the Table.

We have changed "zoonotic"/"zoological" to "host-associated."

It would be also very interesting if the authors would discuss how many of the totally number of OTUs, they generate from the SILVA data remain unaffiliated with any of the known phyla, if any.

All of the bacterial sequences we obtained from the SILVA database affiliated with a phylum (Supplementary Table 6). Among the archaeal sequences two OTUs did not affiliate with a phylum (Supplementary Table 7).

Reviewer #2 (Comments for the Author):

Overall I believe this manuscript is an important and useful extension of the previous microbial census paper. I think the results will be

important for the entire community of microbial ecologists and also many others.

I have some minor concerns with various statements in the manuscript. They are below

We appreciate the reviewer's support for our manuscript. We have done our best to incorporate their comments into the new version of the manuscript.

Line 2: A census is typically carried out for people at a national level. This is simply not true - censuses are carried out for states, counties, cities, towns, and more

Agreed. We have changed this to point out that a census is typically done "across a range of geographic levels."

Line 6: it would be good in the abstract to mention why only full length sequences are being considered for this census

We have added a sentence to the abstract.

Line 21: I think it would be worth addressing the recent issues in regard to whether the archaea are in fact a separate domain from eukaryotes

This is certainly an interesting question; however, given our clarification that we are only concerned with archaeal and bacterial 16S rRNA gene sequences, we feel that this debate is not entirely pertinent to the overall message of the current manuscript.

Line 24: again I think it would be important to explain here why the focus on full length sequences?

Given the lack of space allowed to us for the Abstract and Importance sections, we have only included this explanation in the Abstract. We also justify this more completely in the main body of the manuscript.

Line 33: I would argue this is fundamental to more than just ecology

?????

Line 42. Regarding the deposition of this sequence in 1983 in Genbank at the NCBI.. I don't think this is possible. I think NCBI did not take over Genbank until later. Plus the paper was in 1978 and the precursor toe Genbank existed before 1983 so it would be good to expain the discrepancies.

We have edited the sentence here and in the Conclusion section to remove emphasis on the dates. Here we point out that it was released with the rest of GenBank in 1983.

Line 45: "making it the best-represented gene." details and or/ reference needed for this sentence

?????

Line 68-69. "The number of OTUs that are sampled when using different regions". Not completely clear what is meant here - do you mean the # that are inferred from the data or the # that are actually sampled by the PCR for the different regions?

We have added a clause to point out the number of OTUs that are inferred from the data. It's also likely true that the PCR sampling is also biased between regions.

Line 74: it would be helpful to clarify what is meant here by "lack the references necessary"

We have edited this sentence to clarify that the necessary references are "full-length reference sequences"

Line 92. I think the only thing that can be said from this is saturation for the types of sites being sampled

We hope this is implied from the analysis and discussion we have provided in the manuscript.

Line 104-105. "it is likely that an even larger number of OTUs have yet to be sampled for both domains." Maybe this is a wording thing but it is unclear to me how we can have "started to saturate" yet have an even larger number of OTUs to be sampled

?????

Line 143-145. Since then there have been 4 new bacterial (CKC4, OC31, S2R-29, and SBYG-2791) and 2 new archaeal candidate phyla (Ancient Archaeal Group and TVG8AR30). Citations for this? and is this classification simply pulled fro Silva or based upon other analysis?

These names were simply pulled from SILVA.

Line 177. "Great concern" seems excessive. I think "concern expressed" would be more appropriate

This has been edited as suggested.

Line 208. "There is concern". Can you clarify - or, perhaps it would be better to say "we are concerned"

This has been edited as suggested.

Line 233. Problems with EMIRGE. I think this is possibly true. But not enough evidence is presented in this paper to make this type of statement. What evidence do you have that any errors are coming from EMIRGE?

It was the high degree of novelty that was observed among the EMIRGE sequences that makes us concerned. Furthermore, considering EMIRGE hasn't been used

to sequence a mock community, it is unclear what the error rates are within contigs generated by the approach.

Line 292 - 322. Much of the text here discusses data from Silva which appears to have mostly come from EMBL. And other text discusses sequences from other sources. This is all fine. But the text as it is written in the main parts of the paper does not provide enough of this for context. I think it would be very beneficial to add a bit more to the main text on the sources of the sequences.

We feel that this section of text adequately describes the sources of data used in the paper: SILVA, IMG, and the individual EMIRGE datatsets. These are rolled out in the Results section as we describe the main analysis (SILVA) and the newer methods (IMG and EMIRGE).

Line 337-338. Figshare. Really good to see what is being released in Figshare and Github for this paper. I would further recommend that sequeneces used in this paper also be included in Figshare so that people do not have to rerun the code in order to get access to the sequences (this is what seems to be required)

We are reluctant to do this since the sequences were not generated by us. In addition, the SILVA sequences are subject to a somewhat odd licensing agreement. The code posted in GitHub describes in detail how to obtain and process the sequences.

Line 363. " grow at a rate too slow to ever reach estimates of 10^6 to 10^11 bacterial species." This needs further details in my opinion. What would it take, for example to get to 10^6? How many years at current rates? What if the rate went up a bit? Then how many years?

We have removed this sentence since it is not directly discussed in the main body of the manuscript.

Line 388 "that were also detected by other." I assume some word is missing here.

We have corrected the text by adding "methods."