Removal of rare sequences from 16S rRNA gene sequence surveys biases the interpretation of community structure data

Patrick D. Schloss†

† To whom correspondence should be addressed:

pschloss@umich.edu

Department of Microbiology and Immunology University of Michigan Ann Arbor, MI 48109

Research article format

Abstract

word choice: sample vs. community

250 words

Importance

150 words

EDIT: Two elements that are held in tension in the analysis of 16S rRNA gene sequence data is how to adequately remove PCR and sequencing artifacts and decrease the granularity of the taxonomic level that is used in the analysis. Previous attempts have included screening for sequencing quality based on quality scores [Kozich/Edgar] followed by a polishing step based on the frequency of the sequences relative to similar sequences [Kozich/Edgar/DeBlur]. Other pipelines model the quality scores and types of errors to cluster sequences directly [Dada2]. But, as a final step many pipelines advocate for removing rare sequences from each dataset prior to outputting the sequence data as amplicon sequence variants (ASVs) [Knight/Edgar/DeBlur/Dada2]. ASVs are often clustered further to generate operational taxonomic units or phylotypes. Some pipelines remove all ASVs that appear once (i.e. singletons) [XXXX], XXXXXX [XXXXXX], XXXXX [XXXXXXX], or XXXXXX [XXXXXX] times prior to further clustering or making ecological comparisions. Notably, the mothur-based pipeline discourages the practice of removing rare sequences.

The abundance-based screening approach assumes that rare ASVs are more likely to be artifacts than more abundant ASVs. Sequencing of mock communities confirms that artifacts tend to be rare. Proponents of abundance-based screening point to their ability to obtain the correct number of ASVs, OTUs, or phylotypes with data generated from sequencing mock communities when rare ASVs are removed. However, this approach effectively overfits the curation pipeline to data generated from a phylogenetically simple community with an atypical community distribution that is often sequenced to a depth that is not achieved with biological samples. It is necessary to think more deeply about the practice of abundance-based screening.

The minimum abundance thresholds that have been proscribed are developed and applied without regard for the total number of sequences generated from each sample. Ignored in their recommendations is the common experience that the number of sequences generated from each sample may vary by two or three orders of magnitude. An ASV that appears once in a sample with 2,000 sequences is more trustworthy than an ASV that appears once in a sample with 100,000 sequences since it has a 50-fold higher relative abundance. But, according to the pipeline recommendations, they are treated as being equally trustworthy. Rather than removing rare ASVs, the approach taken by the mothur pipeline applies the classical ecological approach of rarefaction. Each sample is rarefied to the same sequencing depth so that the number of artifacts that appears in each sample is controlled.

Experience sequencing biological samples demonstrates that there are good ASVs that may have an abundance below the proscribed threshold. For example, the abundance of an ASV may be below the threshold in some samples or time points and above the threshold in others. However, rarity, both in terms of prevalence and incidence, is an important ecological concept. Removing rare ASVs likely hinders one's ability to ability to make inferences about the dynamics and nature of the populations that rare ASVs represent.

Furthermore, removing ASVs whose abundances are below the proscribed threshold also potentially biases the community structure of the samples.

In the current study, I use published sequence data from 12 studies to investigate the nature of rare ASVs (i.e. those that appear 10 or fewer times) and the effects of removing them on downstream analysis of microbial communities. The analysis was also performed using traditional operational taxonomic units, where ASVs subjected to abundance-based screening were clustered such that the ASVs within an OTU were no more than 3% different from each other. The results reject the assumptions built into abundance-based screening and highlight the problems inherent in removing rare ASVs.

Results

Datasets. I collected 12 publicly available datasets that used the Illumina MiSeq platform to sequence the V4 region of the 16S rRNA gene from a variety of environments (Table 1). To insure the highest possible data quality, datasets were limited to those where the 500 cycle v2 MiSeq chemistry was used to sequence the amplicons. The paired 250 nt reads resulted in near complete 2-fold sequencing coverage of every nucleotide in the ca. 250 nt-long region. This region and sequencing platform were selected because previous work has shown that a standard data analysis pipeline in mothur results in a sequencing error rate below 0.02% (1). All sequence data were obtained from the Sequence Read Archive and processed using a standard mothur-based sequencing pipeline that resulted in ASVs as generated by the pre.cluster algorithm (1, 2). After removing poor quality and chimeric ASVs and samples that had uncharacteristically low number of sequences for the dataset, these datasets included between 7 and 490 samples (Figure S1). The median number of sequences for each dataset ranged between 6,477 and 193,464 (Table 1). Strikingly, aside from the relatively small marine and soil datasets, the difference between the sample with the fewest sequences and the sample with the most sequences for each dataset varied by between 7.4 and 96.6-fold (Table 1).

The nature of singletons. Removal of rare ASVs is commonly justified as a method of removing ASVs that are artifacts. If such ASVs are artifacts, then one would expect the number of singleton ASVs to accumulate with sequencing depth. Contrary to this expectation, the median percentage of sequences that were discarded when singleton ASVs were removed from each dataset varied between 0.42 and 22.23% (bioethanol and seagrass). In addition, with the exception of the samples from the marine and sediment datasets (Spearman correlation, $P>0.05$), the fraction of singleton ASVs in samples was negatively correlated with the number of sequences in each sample with a range between -0.27 and -0.87 (rice and

bioethanol) (Figure 1A). This showed that with additional sequencing, the probability of seeing singleton ASVs in multiple samples was greater than the probability of generating an artifact. This suggests that the singleton ASVs are not as likely to be artifacts as previously thought. Furthermore, if singleton ASVs were artifacts, then one would not expect to find them in other samples from the same dataset. In fact, singleton ASVs from samples with fewer sequences were often found in samples with more sequences. At least 50% of the singleton ASVs found in the samples from the mice, rice, seagrass, and stream datasets were found in another sample from the same dataset (Figure 1B). Considering the likelihood of finding an ASV duplicated in another sample is confounded by the number of samples and inter-sample diversity, the high coverage of singleton ASVs in these datasets was remarkable. The correlation between the number of sequences in a sample and the fraction of that sample's singleton ASVs that were covered by another sample in the dataset was significant and negative for 9 of the datasets ranging between -0.31 and -0.84 for the rice and seagrass datasets, respectively (Figure 1C). The negative correlation indicated that the singleton ASVs in the smaller samples were more likely to be covered by ASVs in the larger samples. Among the three datasets without a significant correlation (Spearman correlation, $P>0.05$), the marine and soil datasets had the fewest samples in our collection and the stream dataset already had a high level of coverage regardless of the number of sequences. Contrary to the common motivation for removing rare ASVs, these results indicate that this practice disproportionately impacts samples with fewer sequences and likely removes more non-artifact ASVs than those that are artifacts.

The impact of removing rare ASVs on the information represented in each sample. Removing rare ASVs will reduce the richness of ASVs (i.e. the number of ASVs per sample) and increase the relative abundance of the remaining ASVs. To quantify the effect of removing rare ASVs on the information contained within each sample, I varied the minimum abundance threshold to simulate removing ASVs of varying rarity from each sample. The richness of ASVs in each sample decreased by between 34.4 and 86.2% (peromyscus and soil) when removing those ASVs that only appeared once and by between 76.0 and 95.6% (sediment and soil) when removing those that appeared ten or fewer times from each sample (Figure 2A). Similarly, the Shannon diversity decreased by between 1.8 and 15.9% (human and soil) when removing ASVs that only appeared once and by between 5.4 and 35.4% (human and seagrass) when removing ASVs that appeared ten or fewer times from each sample (Figure 2B). Next, I assigned the ASVs to OTUs, which were defined as a group of ASVs that were more than 97% similar to each other to assess the impact of removing rare ASVs on higher level taxonomic groupings that are commonly used in microbial ecology studies. Although pooling similar ASVs into OTUs reduced the impact of removing the rare ASVs relative to the ASV-based analysis, the minimum abundance threshold still decreased the richness of OTUs and

the diversity decreased relative to the full community (Figure S2AB). In contrast to the richness and diversity measurements, the Kullback–Leibler divergence compares the relative abundance of specific ASVs or OTUs between representations of the community. I calculated the Kullback–Leibler divergence between the full communities and those where rare ASVs were removed. As the threshold for removing ASVs increased, the amount of information lost also increased for both ASVs and OTUs (Figure 2C and Figure S2C). The relative loss of information was generally smaller for OTUs than than it was for ASVs. Removing rare ASVs, regardless of abundance threshold, had profound impacts on the representation of the communities.

Removing treatment group effects from community data. Because treatment effects often affect a sample's diversity and inter-sample variation, I generated null distributions for each study by randomizing, without replacement, the number of times each ASV was observed in each sample such that the total number of sequences in each sample and the total number of times each ASV was observed across all samples in the study was the same as was originally observed. This effectively made every community in a study a statistical sample of the study-wide composite community distribution. For example, after this procedure, 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. Because of the risk of bias if only one representation of the null distribution was generated, I generated 100 randomized datasets for each study. The trends between removing rare ASVs and the richness, diversity, and information loss that were identified using the observed community community distribution data were also identified with the data from the null distribution; however, the losses were larger when using the null distribution data (Figure S3). The null distribution data were used in the remainder of the study to minimize the risk of bias.

The impact of removing rare ASVs on the information represented between samples. Considering the loss of richness, diversity, and information when a community has its rarest ASVs removed, it seemed likely that the relationship between communities would also be altered. To assess the impact of removing rare ASVs on measures of alpha diversity between samples I calculated the coefficients of variation (COVs, i.e. the standard deviation divided by the mean) for richness and diversity for each study at multiple abundance thresholds. The COVs for the richness of ASVs across the studies after removing singletons were between 3.6 and 32.7-times larger than they were without removing singleton ASVs (mice and stream; Figure 3A). Similarly, the COVs for the diversity of ASVs were between 1.8 and 20.4-times larger when singletons were removed than when they were not removed (mice and rice; Figure 3B). To assess the impact of removing rare ASVs on measures of beta diversity between samples, I calculated the COVs of the Bray-Curtis distances between samples within the same study at multiple abundance thresholds. The COVs

between Bray-Curtis distances within a study when singletons were removed was between 1.3 and 18.6-times larger than when they were not removed (mice and stream; Figure 3C). When ASVs were clustered into OTUs the difference in COVs was less than it was for the ASVs (Figure S4). These results indicate that removing rare ASVs increases the dissimilarity between samples, which could have a significant impact on the statistical power to detect differences between treatment groups.

The impact of removing rare ASVs on the ability to detect statistically significant differences between treatment groups. To test the effect of increased inter-sample variation, I randomly assigned samples to one of two treatment groups. In the first treatment group, communities were randomly sampled from the null distribution as described above. For the second treatment group, I randomly selected 10% of the ASVs in the pooled study distribution to increase their abundance by 5%. I randomly generated 100 simulated sets of treatment groups and samples. I then tested the ability to detect a difference between the two treatment groups using alpha and beta diversity metrics. The fraction of significant tests was a measurement of the statistical power to detect the difference between the treatment groups. When considering the differences in richness and diversity, the marine dataset yielded no simulated sets that were statistically significant, which was likely due to the small number of samples in the study (N=7). Among the remaining datasets, the power to detect a difference in the richness of ASVs ranged between 0.10 and 0.49 (sediment and stream) and between 0.10 and 0.53 (rainforest and stream) to detect a difference in diversity when using a Wilcox test (Figure 4A). When singleton ASVs were removed, the power to detect a difference in the diversity of ASVs dropped by between 27.3 and 92.9% (bioethanol and soil) and by between 40.0 and 93.3% (rainforest and soil; Figure 4B). The effect of removing rare ASVs on the richness of OTUs and their diversity was similar (Figure S5AB). I used the Bray-Curtis dissimilarity index to compare the simulated communities within each dataset and calculated the power to detect differences between the two simulated treatment groups using the analysis of molecular variance (also called PERMANOVA) (Figure 4C and S5C). Without removing rare sequences, the power to detect a difference between the two simulated treatment groups varied between 0.41 and 1.00 (rainforest and rainforest). Aside from the bioethanol, human, and mice datasets, the power to detect differences dropped by between 6.5 and 64.0% (soil and rice) when singletons were removed. However, when ASVs that occurred 10 or fewer times were removed from each sample, the power to detect differences dropped by 12.0 and 97.2% (human and peromyscus); similar results were observed when ASVs were clustered into OTUs. Removing rare ASVs reduced the ability to detect simulated treatment effects using metrics commonly used to compare microbial communities.

The impact of removing rare ASVs on the probability of falsely detecting a difference between treatment groups. I next asked whether removing rare ASVs could lead to falsely claiming that a treatment

effect had a significant effect on community diversity and structure. 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. Testing at an experiment-wise error rate of 0.05, I expected 5% of the iterations for each dataset to yield a significant test result. Indeed, there was no evidence that removing rare ASVs resulted in an inflated experiment-wise error rate. The average fraction of significant tests did not meaningfully vary from 0.05 across the minimum abundance threshold, dataset, metric of describing sample alpha-diversity, or whether the abundance of ASVs or OTUs were used (Figure 5A and S6A). Similarly, the average fraction of significant tests did not meaningfully vary from 0.05 when using analysis of molecular variance to compare communities using Bray-Curtis distances (Figure 5A and S6A). Second, I again sampled sequences from the null distribution, but assigned samples to one of two treatment groups based on the number of sequences in each sample. The samples with fewer than the median number of sequences for the dataset were assigned to one group and those with more than the median were assigned to the other. This exaggerated bias has been observed in comparisons of the lung and oral microbiota because of the larger number of non-specific amplicons that can be sequenced from lung samples relative to those in the oral cavity leading to a significant difference in sequencing depth between treatment groups []. When rare sequences were not removed, the fraction of significant tests did not differ from 5% for comparing the richness, their diversity, or Bray-Curtis distances (Figure 5B and S6B). However, when rare taxa of any frequency were removed, the probability of falsely detecing a difference as signifiant increased with the definition of rarity (Figure 5B and S6B). Not including the small marine dataset, the average fraction the average fraction of falsely detecting a difference across datasets when only singletons were removed was 92.45%. If there is any relationship between the number of sequences and the treatment group, the risk of falsely rejecting the null hypothesis is inflated when researchers use the strategy of removing rare sequences. The most conservative approach is to not remove low abundance sequences.

Discussion

Removing rare sequences decreases the diversity represented by 16S rRNA gene sequence data and increases the variation between samples. Such impacts will hinder the statistical power to differentiate between treatment groups. Instead of removing rare sequences, researchers should focus on optimizing their sequence generation to minimize the amount of PCR and sequencing errors. In addition, samples should be rarefied to a common number of sequences across samples without prior culling of rare sequences. The number of artifacts is correlated to the number of sequences being considered. With this in mind, rarefaction allows one to control for uneven sampling effort and to control for the number of artifacts in the analysis.

Need to treat rare sequences with a grain of salt

Materials and Methods

- sequencing pipeline description

Acknowledgements

Table 1. Summary of studies used in the analysis. For all studies, the number of sequences used from each study was rarefied to the smallest sample size. A graphical represenation of the distribution of sample sizes for each study and the samples that were removed from each study are provided in Figure S1.

Study	Samples	Total sequences	Median sequences	Range of sequences	Fold-difference between largest and smallest sample
Bioethanol	95	3,972,943	16,015	3,688-356,136	96.6
Human	490	20,909,768	32,505	10,523-430,415	40.9
Lake	52	3,169,868	69,041	15,347-112,871	7.4
Marine	7	1,391,396	193,464	133,516-254,060	1.9
Mice	348	2,813,747	6,477	1,804-30,565	16.9
Peromyscus	111	1,555,545	12,446	4,464-33,644	7.5
Rainforest	69	946,295	11,561	4,932-37,767	7.7
Rice	490	22,591,168	43,216	2,776-193,464	69.7
Seagrass	286	4,130,454	13,567	1,803-45,191	25.1
Sediment	58	1,154,174	17,584	7,685-68,321	8.9
Soil	18	956,656	51,844	47,806-59,956	1.3
Stream	201	21,162,574	90,159	9,175-390,964	42.6

Figure 1.

Figure 2.

References

1. **Kozich JJ**, **Westcott SL**, **Baxter NT**, **Highlander SK**, **Schloss PD**. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and environmental microbiology **79**:5112–5120.

2. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541. doi:10.1128/aem.01541-09.