word choice: sample vs. community

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

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 phylogypes 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% (**???**). 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). After removing poor quality and chimeric ASVs and samples that had uncharacteristically low number of sequences for the dataset, these datasets included between 7.00 and 490.00 samples. The median number of sequences for each dataset ranged between 6,477.00 and 193,464.00 (Table 1, Figure S1). 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 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 the sample was negatively correlated with the number of sequences in each sample with a range between -0.27 and -0.87 (rice and bioethanol). This showed that with additional sequencing, the probability of seeing rare ASVs multiple times was greater than the probability of generating an artifact. This suggests that the rare ASVs are not as likely to be artifacts as previously thought. Furthermore, if rare 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. 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. 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 a large number of non-artifact ASVs.

How do communities change when rare ASVs are removed? Removing rare ASVs will reduce the number of observed ASVs and proportionally increase the relative abundance of the remaining ASVs. The result was expected to be a loss of information contained within each sample. 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 number of ASVs in each sample decreased by between 34.4 and 86.2% when removing those ASVs that only appeared once and by between 76.0 and 95.6% when removing those that appeared ten or fewer times from each sample (Figure 1). Similarly, the Shannon diversity decreased by between 1.8 and 15.9% when removing ASVs that only appeared once and by between 76.0 and 95.6% when removing ASVs that appeared ten or fewer times from each sample (Figure 1). Next, I assigned the ASVs to OTUs 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 number of observed OTUs and the diversity decreased relative to the original community (**Figure 1**). In contrast to the number of ASVs or OTUs and their diversity, the Kullback–Leibler divergence compares the relative abundance of each ASV or OTU between representations of the community. I calculated the Kullback–Leibler divergence from the full to pruned communities when rare ASVs were removed. As the threshold for removing ASVs increased, the amount of information lost also increased for both ASVs and OTUs (Figure 1). The relative loss of information was generally lower 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 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 number of ASVs and diversity 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 same trends between removing rare ASVs and the number of ASVs, sample diversity, and information loss that were identified using the observed community community distribution data were also identified with the data from the null distribution (**Figure S2**). The null distribution data were used in the remainder of the study.

How do the relationships between communities change when they are screened by removing rare ASVs. Considering the loss of the number of ASVs, diversity, and information when a community has its rarest ASVs removed, it seemed likely that the relationship between communities would be altered. To assess the impact of removing rare ASVs on measures of alpha diversity between samples I calculated the coefficients of variation (COV, i.e. standard deviation divided by the mean) for the number of ASVs and diversity for each study at multiple abundance thresholds. The COV for the number of ASVs across the studies after removing singletons were between 3.6 and 32.7-times larger than they were without removing singleton ASVs. 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.To assess the impact of removing rare ASVs on measures of beta diversity between samples I calculated the COV of the Bray-Curtis distances between samples within the same study at multiple abundance thresholds. The COV 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. Increasing the minimum abundance threshold increased the COV between samples when using metrics of alpha and beta diversity. When ASVs were clustered into OTUs the difference in the COV was less than it was for the ASVs (**Figure 2**). 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.

How does the increased inter-sample variation impact the ability to detect differences between samples? 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 increased the abundance of 10% of the ASVs in the pooled study distribution 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. Across all of the metrics I tested, 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); the marine dataset is not included in the following analysis. Among the remaining datasets, the power to detect a difference in the observed number of ASVs ranged between 0.10 and 0.49 and between 0.10 and 0.53 to detect a difference in diversity when using a Wilcox test (**Figure X**). When singleton ASVs were removed, the power to detect a difference in the number of ASVs dropped by between 27.3 and 92.9% and by between 40.0 and 93.3% to detect a difference in diversity. The effect of removing rare ASVs on the number of OTUs and their diversity was similar. A popular alternative approach calculates the dissimilarity between communities based on the abundance of each ASV or OTU. We 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). Without removing rare sequences, the power to detect a difference between the two simulated treatment groups varied between 0.41 and 1.00. Aside from 3 datasets, the power to detect differences dropped by between 6.5 and 64.0% 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%; 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.

How does the removing of rare ASVs impact the probability of falsely detecting a difference between treatment groups? Observing reduced ability to detect differences between communities when rare ASVs were removed from each sample, 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, for each dataset, I sampled sequences from the null distribution and randomly assigned each sample to one of two treatment groups and determined the number of observed ASVs and OTUs and their Shannon diversity index. 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 XA**). 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. 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 nubmer of non-specific amplicons that can be sequenced from lung samples relative to those in the oral cavity [**REF**]. When rare sequences were not removed, the fraction of significant tests did not differ from 5% for comparing the number of observed taxa (using ASVs or OTUs), their shannon diversity, or Bray-Curtis distances. 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 XB**). Across the datasets, the average fraction the average fraction of falsely detecting a difference was when only singletons were removed was greater than 90%. 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.

Conclusion. 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

Acknowledgements

Materials and Methods

- sequencing pipeline description

References

Figure 1.

Figure 2.

1. **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.