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

Patrick D. Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

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

Observation format

Abstract

Importance

Question: What impact does removing singletons (and other rare sequences) have on downstream data analysis?

- Many pipelines remove rare sequences prior to clustering sequences into OTUs or as part of the generation fo ASVs
- Singletons can be sequences that appear 1-10 times in larger samples
- Seeks to solve the problem of low quality sequence data - more singletons with higher error rates
- Likely there are rare sequences that are good - e.g. high diversity samples
- People have overfit their pipelines ot mock community data which have unrealistic community distributions (i.e. uniform abundance or log abundance) and phylogenetic diversity
- Better solutions would include:
  - Get better sequence data
  - Rarefy data so that the number of errors are controlled for across samples
- Not all samples are sequenced evenly, so a singleton doesn't hav the same relative abundance across all samples

H: Removing rare sequences from each sample will skew the community distribution to impact the amount of variation in the relative abundance and alpha and beta diversity values across samples

Acknowledgements

Materials and Methods

References

Figure 1.

Figure 2.