Amplicon sequence variants should not replace operational taxonomic units in marker-gene data analysis

Running title: ASVs vs. OTUs

Patrick D. Schloss†

† To whom corresponsdence should be addressed:

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Observation Format

11 **Abstract (250 words)**

12 **Importance (150 words)**

16S rRNA gene sequencing is a very powerful technique for describing and comparing microbial communities. Efforts to link 16S rRNA gene sequences to taxonomic levels based on distance thresholds go back to at least the 1990s. The distance-based thresholds that were developed and are now widely used (3%) were based on DNA-DNA hybridization approaches that are not as precise as genome sequencing. Instead, genome sequencing technologies have suggested that the widely used 3% distance threshold to operationally define bacterial taxa is too coarse. As an alternative to OTUs, amplicon sequencing variants (ASVs) have been proposed as a way to adopt the thresholds suggested by genome sequencing to microbial community analysis using 16S rRNA gene sequences. ASVs are a unit of microbial community inference that do not cluster sequences based on a distance-based threshold. However, most bacterial genomes have more than 1 copy of the rrn operon and those copies are not identical. Therefore, using too fine a threshold to identify OTUs creates the risk of splitting a single genome into multiple bins and using too broad of a threshold to define OTUs creates the risk of lumping together bacterial species into the same OTU. An example of both is seen in the comparison of *Staphylococcus aureus* (NCTC 8325) and *S. epidermidis* (ATCC 12228) where each genome has 5 copies of the 16S rRNA gene. The 10 copies of the 16S rRNA gene each have a different sequence and so if OTUs are defined based on ASVs, each genome would be split into 5 OTUs. Conversely, if the copies were clustered using a 3% distance threshold all 10 copies would cluster into the same OTU. The goal of this study was to quantify the risk of splitting a single genome into multiple bins and the risk of lumping together different bacterial species into the same bin.

To investigate the variation in the number of copies of the 16S rRNA gene per genome as well as the intragenomic variation among copies of the 16S rRNA gene, I obtained reference 16S rRNA sequences from the rrn copy number database (rrnDB; CITATION). Among the **4,774** species represented in the rrnDB there were **15,614** genomes. The median number of rrn operson per species ranged between **1** (e.g. ***Mycobacterium tuberculosis***) and **19** (***Metabacillus litoralis***) copies of the rrn operon. As the number of copies of the operon in a genome increased, the number of variants of the 16S rRNA gene in each genome also increased (**FIGURE**). On average, there were **1** variants per copy of the full length 16S rRNA gene and an average of **0.26**, **0.33**, and **0.27** variants when considering the V4, V3-V4, and V4-V5 regions of the gene, respectively. Although a species tended to have a consistent number of 16S rRNA gene copies per genome, the number of total variants increased with the number of genomes that were sampled (**FIGURE**). For example, ***Mycobacterium tuberculosis*** generally only had **1** copy of the gene per genome, but across the **180** genomes that have been sequenced there were **11** versions of the gene. Similarly, a *E. coli* genome typically had **7** copies of the 16S rRNA gene with between **6** and **10** distinct full length sequences per genome. Across the **958** *E. coli* genomes that have been sequenced, there were **1,013** different variants of

the gene. These observations highlight the risk of selecting a threshold for defining units of inference that is too narrow because it is possible to split a single genome into multiple units.

A method to avoid splitting a single genome into multiple units of inference is to cluster 16S rRNA gene sequences together that are similar. However, this also increases the risk of lumping together genes from different species that are similar to each other. Therefore, I assessed the impact of the threshold used to define clusters of 16S rRNA genes on the propopensity to lump species together and split genome apart. I identified the threshold where 90% of bacterial species would be represented by a single OTU. For full length 16S rRNA gene sequences, I found that at a threshold of XX%, 90% of the species would be represented by a single OTU. Similarly, thresholds of XX, XX, and XX% were observed for the V4, V3-V4, and V4-V5 regions. However, at these thresholds, multiple species could be represented by the same OTU. At the highest level of resolution, XX% of the species shared a 16S rRNA gene sequence variant with another species. Given the risk of splitting a genome into multiple OTUs is more biologically problematic than lumping species together, larger thresholds are advisable.

To provide a more nuanced approach to selecting a threshold, it would be useful to to quantify the sensitivity and specificity of characterizing bacterial species using OTUs defined at different thresholds. I created confusion matrices for multiple regions of the 16S rRNA gene: true positives were those cases where two ASVs were joined in the same OTU and the same species; true negatives were those cases where two ASVs from different OTUs came from different species; false positives were those ASVs that joined the same OTU, but were from different species; and false negatives were those ASVs that joined different OTUs, but were from the same species. By calculating the sensitivity and specificity for each threshold and each region of the 16S rRNA gene, I was able to constuct a receiver operator characteristic curve (ROC). Because the ROC curve represents a range of possible thresholds and sensitivities and specificities, I used two metrics to select the best threshold for defining an OTU. First, I identified the thresholds where the sensitivity and specificity were most similar to each other. For this criterion, the best distance thresholds were **6.0**% (V1-V9), **4.5**% (V4), **5.5**% (V3-V4), and **4.0**% (V4-V5). Second, I identified the distance threshold that resulted in the point on the ROC curve that was closest to perfect classification. For this criterion, the best distance thresholds were **5.5**% (V1-V9), **3.5**% (V4), **4.5**% (V3-V4), and **3.5**% (V4-V5). Surprisingly, these analyses revealed that thresholds near 3% distance balance the risks of splitting genomes into separate OTUs and lumping species into the same OTU.

The results of this analysis demonstrate that there is a significant risk of splitting single genomes into multiple bins if too fine of a threshold is applied to defining an OTU. An ongoing problem for amplicon-based studies is defining a meaningful taxonomic unit of inference. Since there is no consensus definition for a biological

species concept, microbiologists must accept that how we have named bacterial species is biased and that taxonomic rules are not applied in a consistent manner. This makes it more challenging to attempt to fit a distance threshold to define an OTU definition that matches a set of species names. Furthermore, it is unlikely that the 16S rRNA gene evolves at the same rate across all bacterial lineages, which limits the biological interpretation of a common OTU definition. At best, a distance-based definition of a taxonomic unit is operational. There is general agreement in bacterial systematics that to classify something to a bacterial species, you need phenotypic and genome sequence data (CITATION). We are asking too much of a short section of a bacterial genome to be able to differentiate between species. It is difficult to defend a unit of inference that would split a single genome into multiple taxonomic units. It is not biologically plausible to entertain the possability that parts of a genome would have different ecologies. Although there are multiple reasons that proponents of ASVs encourage their use, the significant risk of splitting genomes is too high to warrant their use.

Materials and Methods. (i) Data availability. The 16S rRNA gene sequences used in this study were obtained from the *rrn*DB (https://rrndb.umms.med.umich.edu; version 5.6, released November 8, 2019). At the time of submission, this is the most current version of the database. The *rrn*DB obtained the curated 16S rRNA gene sequences from the KEGG database, which ultimately obtained them from NCBI's non-redundant RefSeq database. The *rrn*DB provides downloadable versions of the sequences with their taxonomy as determined using the naive Bayesian classifier trained on the RDP reference taxonomy. For some genomes this resulted in multiple classifications since a genome's 16S rRNA gene sequences were not identical. Instead, I mapped the RefSeq accession number for each genome in the database to obtain a single taxonomy for each genome. Because strain names were not consistently given to genomes across bacterial species, the strain level designations were ignored.

(ii) Definition of regions within 16S rRNA gene. The full length 16S rRNA gene sequences were aligned to a SILVA reference alignment of the 16S rRNA gene (v138) using the mothur software package (v. 1.XX). Regions of the 16S rRNA gene were selected because of their use in the microbial ecology literature. Full length sequences corresponded to *E. coli* positions XX through XXXX, V4 to positions XXX through XXX, V3-V4 to positions XXX through XXX, and V4-V5 to positions XXX through XXX.

(iii) Controlling for uneven sampling of genomes by species. Because of the uneven distribution of genome sequences across species, for the analysis of splitting genomes and lumping species I randomly selected one genome for each species. The random selection was repeated 100 times. Analyses based on this randomization report the median of the 100 randomizations. The intraquartile range between randomizations was typically less than XXXX. Because it was so small, confidence intervals are not included

in Figure 2.

(iv) Reproducible data analysis. The code to perform the analysis in this manuscript and its hisotry are available as a git-based version control repository on GitHub (https://github.com/pschloss/Schloss_rrnAnalysis_XXXX_2020). The analysis can be regenerated using a GNU Make-based workflow that made use of built-in bash tools (v. 3.2.57), mothur (v. 1.XX), and R (v. 4.X.X). Within R, I used the tidyverse (v. 4.X.X), data.table (v. 4.X.X), Rcpp (v. 4.X.X), furrr (v. 4.X.X), and rmarkdown (v. 4.X.X) packages. The conception and development of this analysis is available as a playlist on the Riffomonas YouTube channel (https://www.youtube.com/playlist?list=PLmNrK_nkqBpKY3SZiivIIGvcLX-KHmfR8).

Acknowledgements. I am grateful to Robert Hein and Thomas Schmidt who maintain the rrnDB for their help in understanding the curation of the database and for making the 16S rRNA gene sequences and related metadata publicly available. I am also grateful to community members who watched the serialized version of this analysis on YouTube and provided their suggestions and questions.

This work was supported, in part, through grants from the NIH to PDS (P30DK034933, U01AI124255, and R01CA215574).

References

Figures