Amplicon sequence variants artificially split bacterial genomes into separate units of inference

Running title: ASVs artificially split bacterial genomes

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

Abstract

Amplicon sequencing variants (ASVs) have been proposed as an alternative to operational taxonomic units (OTUs) for analyzing microbiomes. ASVs have grown in popularity, in part, because of a desire to reflect a more refined level of taxonomy because they do not cluster sequences based on a distance-based threshold. However, ASVs and the use of overly narrow thresholds to identify OTUs increases the risk of splitting a single genome into separate clusters. I addressed this problem by analyzing the intragenomic variation of 16S rRNA genes from the bacterial genomes represented in the rrn copy number database, which contained **15,614** genomes from **4,774** species. The analysis confirmed the severity of this risk. As the number of copies of the 16S rRNA gene increased in a genome, the number of ASVs also increased. For full length 16S rRNA genes, there was an average of **0.60** ASVs per copy of the 16S rRNA gene. Among genomes with 7 copies of the 16S rRNA, such as *E. coli*, it was necessary to use a distance threshold of **5.5**% to cluster the ASVs from the same genome into a single OTU. This research highlights the risk of splitting a single bacterial genome into separate clusters when ASVs are used to analyze 16S rRNA gene sequence data. Although there is also a risk of clustering different species into the same OTU, the effects of those risks on biological inferences are less than those from artificially splitting a genome into separate ASVs and OTUs.

Importance

16S rRNA gene sequencing has propelled significant interest into host-associated and environmental microbiomes. There is a tension between trying to classify 16S rRNA gene sequences to increasingly lower taxonomic levels and the reality that those levels were defined using more sequence and physiological information than is available from a fragment of the 16S rRNA gene. Furthermore, naming of bacterial taxa reflects the biases of those who name them. One motivation for the recent push to adopt ASVs in place of OTUs in microbiome analyses is to allow researchers to perform their analyes at the finest possible level that reflects species-level taxonomy. The current research is significant because it quantifies the risk of artificially splitting bacterial genomes into separate clusters. Far from providing a better represenation of microbiomes, the ASV approach could lead to conflicting inferences about the ecology of different clusters from the same genome.

16S rRNA gene sequencing is a powerful technique for describing and comparing microbial communities (1). 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 (2, 3). Instead, genome sequencing technologies have suggested that the widely used 3% distance threshold to operationally define bacterial taxa is too coarse (4–6). 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 (7–10). ASVs are a unit of microbial community inference that do not cluster sequences based on a distance-based threshold (11). However, most bacterial genomes have more than 1 copy of the rrn operon and those copies are not identical (12, 13). 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)(14). 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. On average, there were **0.60** 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 1**). 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 together similar 16S rRNA gene sequences. Therefore, I assessed the impact of the distance threshold used to define clusters of 16S rRNA genes on the propensity to split a genome into separate clusters. I observed that as the number of copies of the *rrn* operon increased, the distance threshold required to reduce the ASVs in each genome to a single OTU increased (Figure 1). Among species with 7 copies of the *rrn* operon (e.g., *E. coli*), I found that a threshold of **5.5**% was required to reduce full length ASVs to a single OTU in 95% of the species. Similarly, thresholds of **2.5**, **4.0**, and **3.5**% were required for the V4, V3-V4, and V4-V5 regions, respectively. But, if a 3% distance threshold was used, then ASVs from genomes containing fewer than **5**, **8**, **6**, and **6** copies of the *rrn* operon would reliably be clustered into a single OTU for ASVs from the V1-V9, V4, V3-V4, and V4-V5 regions, respectively. Consequently, these results demonstrate that broad thresholds must be used to avoid splitting different operons from the same genome into separate clusters.

At broad thresholds multiple species could be represented by the same OTU (**Figure 2**). Using ASVs, **3.6**% of the species shared a 16S rRNA gene sequence variant with another species when considering full length sequences and **14.9**, **10.2**, and **12.0**% when considering the V4, V3-V4, and V4-V5 regions, respectively. At the commonly used 3% threshold, **25.2**% of the species shared an OTU when considering full length sequences and **33.0**, **29.4**, and **32.2**% when considering the V4, V3-V4, and V4-V5 regions, respectively. Considering that species designations are unevenly applied and reflect multiple biases, the risk of splitting a genome into multiple OTUs more problematic than clustering species together. Therefore, larger thresholds are advisable.

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 (11, 15, 16). Since there is no consensus definition for a biological species concept (17, 18), microbiologists must accept that how we have named bacterial species is biased and that taxonomic rules are not applied in a consistent manner (e.g., (19)). This makes it more challenging to attempt to fit a distance threshold to define an OTU definition that matches a set of species names (20). Furthermore, the 16S rRNA gene does not evolve at the same rate across all bacterial lineages (15), which limits the biological interpretation of a common OTU definition. A distance-based definition of a taxonomic unit based on 16S rRNA gene or full genome sequences is, at best, operational and not grounded in biological theory (15, 21–23). There is general agreement in bacterial systematics that to classify

something to a bacterial species, you need phenotypic and genome sequence data (17–19). 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 possibility 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) (14). 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) (24, 25). 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

1. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences 82:6955–6959 https://doi.org/10.1073/pnas.82.20.6955.

2. Stackebrandt E, Goebel BM. 1994. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 44:846–849 https://doi.org/10.1099/00207713-44-4-846.

3. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology 57:81–91 https://doi.org/10.1099/ijs.0.64483-0.

4. Rodriguez-R LM, Castro JC, Kyrpides NC, Cole JR, Tiedje JM, Konstantinidis KT. 2018. How much do rRNA gene surveys underestimate extant bacterial diversity? Applied and Environmental Microbiology 84:e00014–18 https://doi.org/10.1128/aem.00014-18.

5. Stackebrandt E, Ebers J. 2006. Taxonomic parameters revisited: Tarnished gold standards. Microbiol Today 33:152–155.

6. Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375 https://doi.org/10.1093/bioinformatics/bty113.

7. Edgar RC. 2016. UNOISE2: Improved error-correction for illumina 16S and its amplicon sequencing. bioRxiv https://doi.org/10.1101/081257.

8. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191–16 https://doi.org/10.1128/mSystems.00191-16.

9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from illumina amplicon data. Nature Methods 13:581–583 https://doi.org/10.1038/nmeth.3869.

10. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2014. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. The ISME Journal 9:968–979 https://doi.org/10.1038/ismej.2014.195.

11. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational

taxonomic units in marker-gene data analysis. The ISME Journal 11:2639–2643 https://doi.org/10.1038/ismej.2017.119.

12. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, DeSantis T, Brodie E, Nelson K, Pei Z. 2010. Diversity of 16S rRNA genes within individual prokaryotic genomes. Applied and Environmental Microbiology 76:3886–3897 https://doi.org/10.1128/aem.02953-09.

13. Sun D-L, Jiang X, Wu QL, Zhou N-Y. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Applied and Environmental Microbiology 79:5962–5969 https://doi.org/10.1128/aem.01282-13.

14. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. 2014. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research 43:D593–D598 https://doi.org/10.1093/nar/gku1201.

15. Schloss PD, Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology 77:3219–3226 https://doi.org/10.1128/aem.02810-10.

16. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature Communications 10:5029 https://doi.org/10.1038/s41467-019-13036-1.

17. Staley JT. 2006. The bacterial species dilemma and the genomicphylogenetic species concept. Philosophical Transactions of the Royal Society B: Biological Sciences 361:1899–1909 https://doi.org/10.1098/rstb.2006.1914.

18. Oren A, Garrity GM. 2013. Then and now: A systematic review of the systematics of prokaryotes in the last 80 years. Antonie van Leeuwenhoek 106:43–56 https://doi.org/10.1007/s10482-013-0084-1.

19. Baltrus DA, McCann HC, Guttman DS. 2016. Evolution, genomics and epidemiology ofPseudomonas syringae. Molecular Plant Pathology 18:152–168 https://doi.org/10.1111/mpp.12506.

20. Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. Journal of Bacteriology 187:6258–6264 https://doi.org/10.1128/jb.187.18.6258-6264.2005.

21. Barco RA, Garrity GM, Scott JJ, Amend JP, Nealson KH, Emerson D. 2020. A genus definition for bacteria and archaea based on a standard genome relatedness index. mBio 11:02475–19 https:

//doi.org/10.1128/mbio.02475-19.

22. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for bacteria and archaea. Nature Biotechnology 38:1079–1086 https://doi.org/10.1038/s41587-020-0501-8.

23. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology 12:635–645 https://doi.org/10.1038/nrmicro3330.

24. 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 https://doi.org/10.1128/aem.01541-09.

25. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41:D590–D596 https://doi.org/10.1093/nar/gks1219.

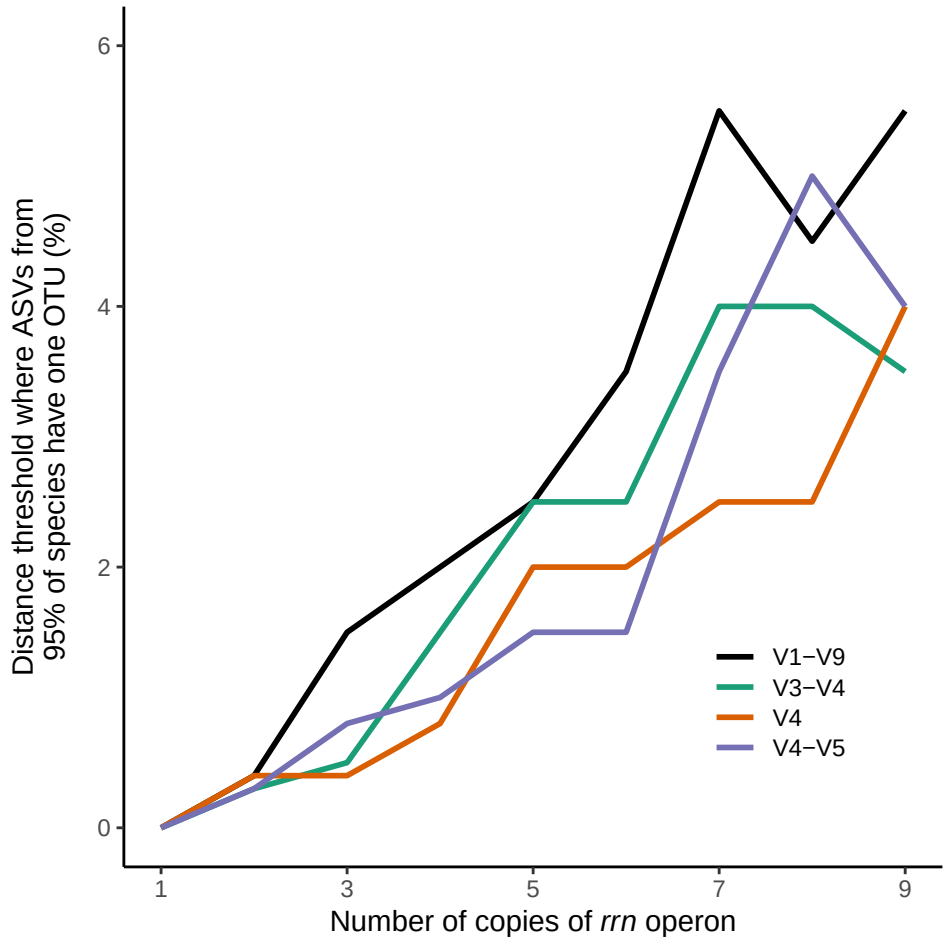
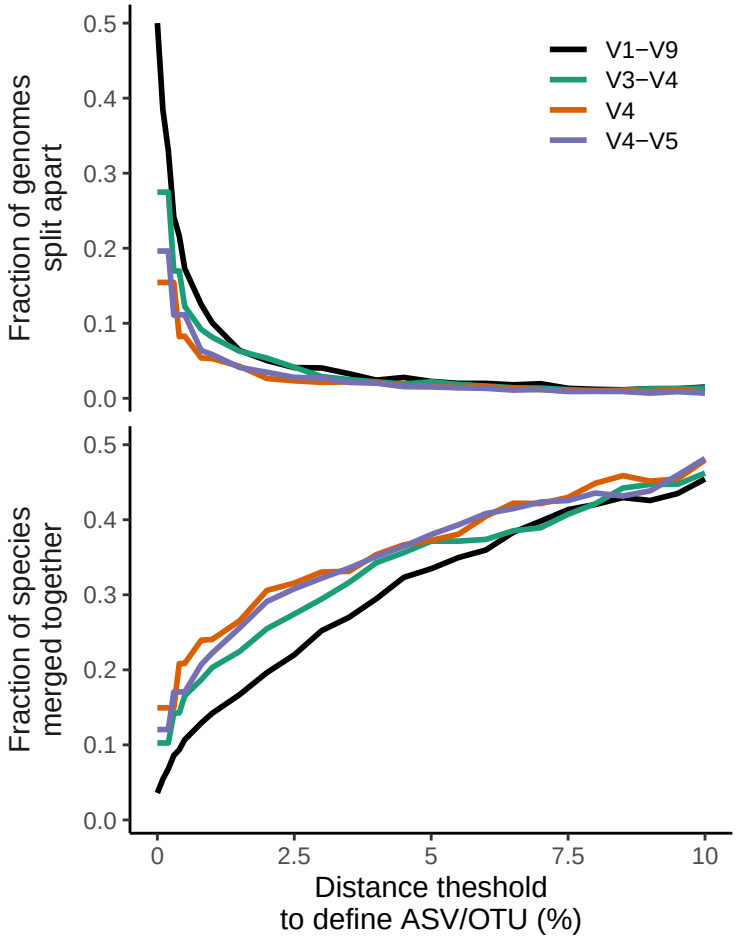


Figure 1. The distance threshold required to prevent the splitting of genomes into multiple OTUs increases as the number of *rrn* operons in the genome increases. Each line represents the median distance threshold for each region of the 16S rRNA gene that is required for 95% of the species with the indicated numbrer of *rrn* operons to cluster their ASVs to a single OTU. The median distance threshold was calculated across 100 randomizations in which one genome was sampled from each species. Only those number of *rrn* operons that were found in more than 100 species are included.



223

224 **Figure 2. As the distance threshold used to define an OTU increases, the fraction of genomes split**
225 **into separate OTUs decreases while the fraction of species that are merged into the same OTU**
226 **increases.** These data represent the median fractions for both measurements across 100 randomizations.
227 In each randomization, one genome was sampled from each species.

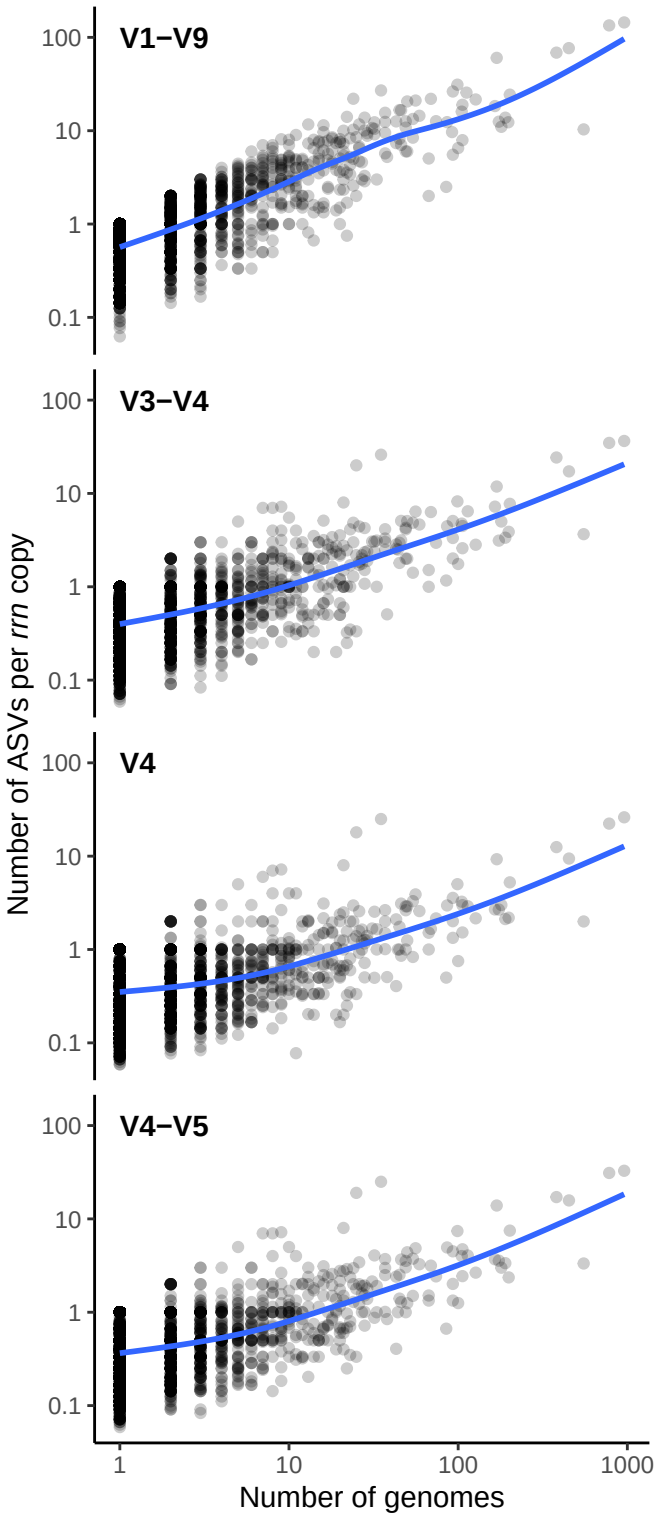


Figure S1. The ratio of number of distinct ASVs per copy of the *rrn* operon increases for a species as the number of genomes sampled increases. Each point represents a different species and is shaded to be 80% transparent so that when points overlap they become darker. The blue line represents a smoothed

fit through the data.