# 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:

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

**Observation Format**

## Abstract (250 words)

## Importance (150 words)

## Introduction

* 16S rRNA gene sequencing is a very powerful technique for describing and comparing microbial communities
* How do we analyze them (classification, clustering)?
* What has changed in recent years? ASVs
* Efforts to link 16S rRNA gene sequences to taxonomic levels based on distance thresholds go back a long way
* ESVs/ASVs have been an attempt to adopt the thresholds suggested by genome sequencing to microbial community analysis using 16S rRNA gene sequences
* Most bacterial genomes have more than 1 copy of the rrn operon and those copies are not identical
* Using too fine a threshold to create taxonomic groups runs risk of splitting single genome into multiple bins
* For example, E. coli K-12 has 7 copies of the 16S rRNA gene with 5 variants
* Using too broad a threshold to define ASVs or OTUs risks lumping together bacterial species into the same grouping
* For example, B. cereus, thuringiensis, anthracis share the same 16S rRNA gene sequences
* Goal of this study

## Results

* ESVs/ASVs
  + copy number varies by taxonomy
  + more copies, more variants per genome
  + full length sequences have more variants than sub-regions
  + as more sequences are added to a species, the number of variants increases
* OTUs
  + increasing a threshold decreases the number of variants
  + this limits the splitting of a single genome into multiple bins
  + this increases the lumping of species into single bin

## Conclusions

* Briefly synthesize results
  + Unlikely that the unit of inference should be an ASV
* No biological argument to split a genome into multiple bins
* This analysis has allowed some splitting to balance with lumping
* To reduce splitting further, you would need larger thresholds
* There is general agreement in the field that if you want to classify something to a bacterial species, you need more than the 16S rRNA gene
* Furthermore, using only a few hundred bases of that gene are even more limited.
* We are asking too much of a short section of sequence
* Surprisingly, 3% performs pretty well for an operational definition that limits splitting of bacterial genomes and avoiding the lumping of bacterial species

## Materials and Methods

* rrnDB
* NCBI taxonomy
* R and R packages
* GitHub / YouTube

## Acknowledgements

## Figures