In silico comparison of the identity between the microbial 16S rRNA gene, operon, and genome

Melanie Tan, Liping Zhao, Guojun Wu

Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA

RUTGERS
Aresty Research Center for Undergraduates

Background

- The **16S rRNA gene** is a commonly used biomarker for microbial genome studies.
- Recent advancement in long-read sequencing techniques, which offer higher resolution than short-read sequencing methodology, have allowed us to further investigate the structure of microbial communities by using **full-length 16S rRNA genes** and **rRNA operons.**
- However, the extent to which the full-length 16S rRNA gene and rRNA operon can capture **genome-level variation** remains elusive.
- Here, we study the relationships between sequence identity across three levels: the **full-length 16S rRNA gene**, the **16S-ITS-23S rRNA operon**, and the **microbial genome**.

Methods

Download 33,406 complete bacterial and archaeal genomes from National Center for Biotechnology Information (NCBI) database

→ Remove genes and operons with abnormal lengths, leaving 31,215 genomes covering 7,089 bacterial and 403 archaeal species

→ Cluster 169,963 16S rRNA genes at a 100% identity using VSEARCH

→ Pairwise comparisons of the gene, operon, and genome using VSEARCH and fastANI

Results

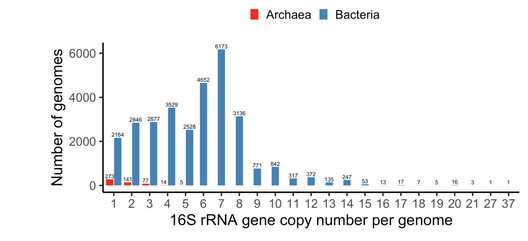

Figure 1. Wide range of 16S rRNA gene copy numbers across archaea and bacteria. Archaeal copy numbers range from 1 to 5 copies. Bacterial copy numbers range from 1 to 37 copies, and genomes with 7 copies were most abundant.

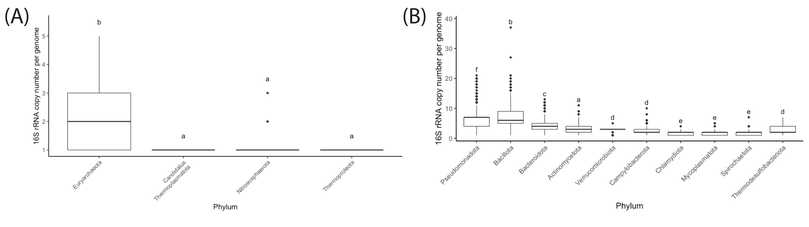

Figure 2. 16S rRNA copy number was taxon specific at the phylum level. (A) Copy number distribution for 4 archaeal phyla with at least 10 genomes. The highest mean copy number belonged to *Euryarchaeota* (1.9 ± 0.9), followed by *Nitrososphaerota* (1.1 ± 0.4). **(B)** Copy number distribution for 10 bacterial phyla with at least 100 genomes. The highest mean copy number belonged to *Bacillota* (*Firmicutes*) at 6.9 ± 2.8 copies, followed by *Pseudomonadota* (*Proteobacteria*) at 6.0 ± 2.3 copies.

Results

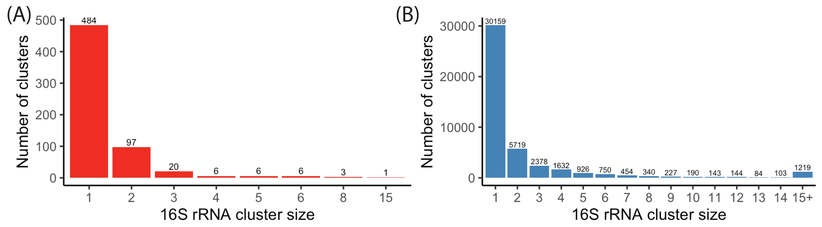

Figure 3. Distribution of operon cluster sizes in archaea and bacterial at a level of 100% identity. (A) 867 archaeal 16S rRNA genes were clustered into 623 clusters. The largest cluster contained 15 genes. **(B)** 169,937 bacterial 16S rRNA genes were clustered into 44,468 clusters. 8 clusters contained over 1,000 genes; the largest cluster contained 2,538 genes that came from 833 genomes that belonged to two species: *Staphylococcus aureus* and *Staphylococcus argenteus.*

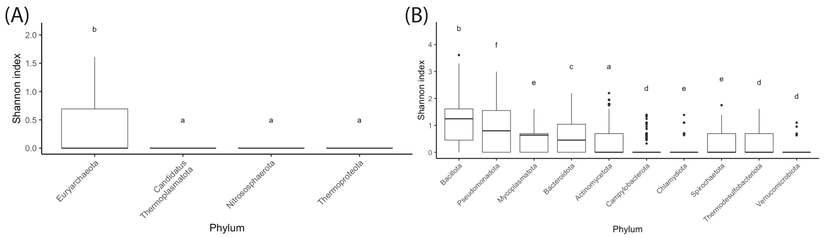

Figure 4. Intragenomic diversity was taxon specific at the phylum level. (A) The Shannon index of diversity was calculated for each genome in archaeal phyla with at least 10 genomes. **(B)** The Shannon index of diversity was calculated for each genome in bacterial phyla with at least 100 genomes.

Table 1. Clustering at a 100% identity leads to over-splitting and over-merging. The over-splitting rate was calculated as the percentage of species split into multiple clusters. The over-merging rate was calculated as the percentage of clusters containing multiple species.

	Archaea	Bacteria	All
Over-splitting (%)	36.48	64.24	62.75
Over-merging (%)	2.25	3.92	3.89

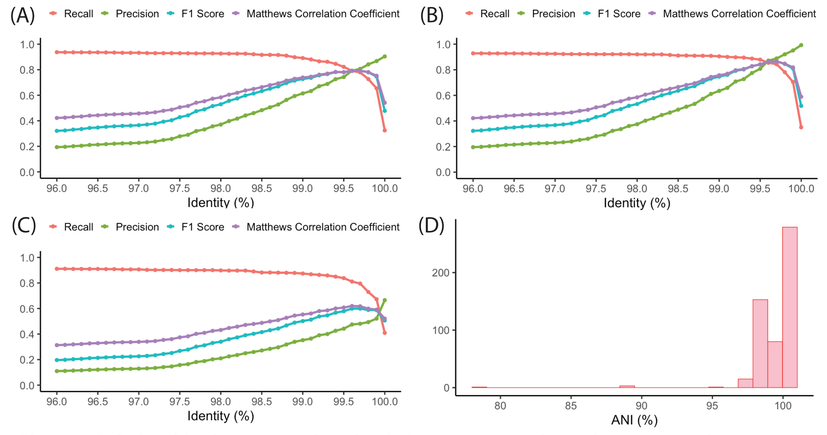

Figure 5. Relationships between archaeal 16S rRNA genes, species assignments, and genomes at identity cutoffs from 96% to 100%. (A) A 99.7% gene identity cutoff best represents species assignment. **(B)** A 99.7% gene identity cutoff best represents genomes that are at least 95% similar. **(C)** A 99.6% gene identity cutoff best represents genomes that are at least 99% similar. **(D)** Pairs of identical archaeal 16S rRNA genes correspond to an average ANI (average nucleotide identity) of 99.29% but can lead to a difference of 22% in some cases.

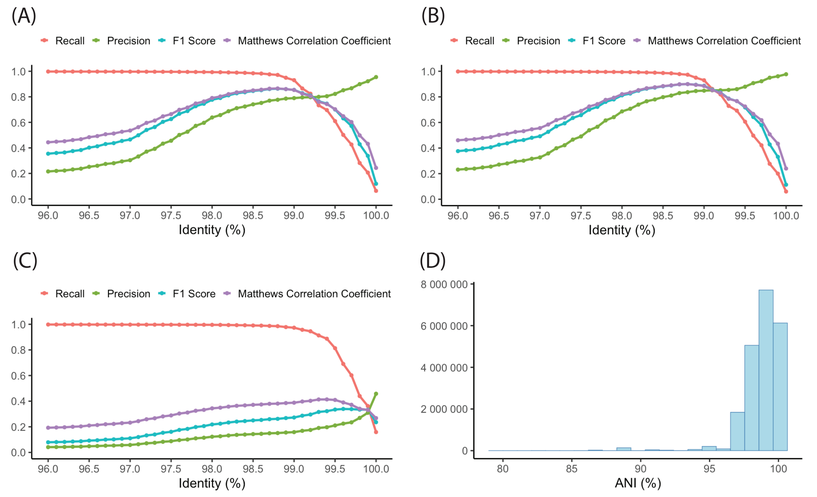

Figure 6. Relationships between bacterial 16S rRNA genes, species assignments, and genomes at identity cutoffs from 96% to 100%. (A) A 98.8% gene identity cutoff best represents species assignment. **(B)** A 98.8% gene identity cutoff best represents genomes that are at least 95% similar. **(C)** A 99.6% gene identity cutoff best represents genomes that are at least 99% similar. **(D)** Pairs of identical archaeal 16S rRNA genes correspond to an average ANI (average nucleotide identity) of 98.71% but can lead to a difference of over 20% in some cases.

Future Directions

- Pairwise comparisons of the 16S-ITS-23S operon and genome will provide additional quantitative data on effective identity cutoffs.

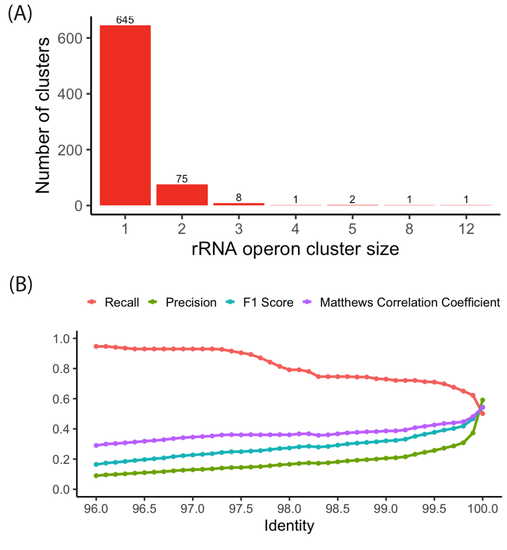

Figure 7. Analysis of archaeal operons. (A) Distribution of operon cluster size in archaea at a level of 100% identity. The largest cluster contained 12 operons from different *Saccharolobus solfataricus* genomes. **(B)** Pairwise comparison shows that a 100% identity cutoff for archaeal 16S-ITS-23S rRNA operons best represents genomes that are at least 99% similar.