Identification of leader-trailer helices of precursor ribosomal RNA in all phyla of bacteria and archaea

Bryan T. Gemler[1,2], Benjamin R. Warner[2, 3], Ralf Bundschuh[1,2,4,5,6, *], Kurt Fredrick[2,3, *]

[1]Interdisciplinary Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210, USA

[2]Center for RNA Biology, The Ohio State University, Columbus, Ohio 43210, USA,

[3]Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA

[4]Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA,

[5]Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, USA,

[6]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, Ohio 43210, USA,

*Correspondence:

Kurt Fredrick: Phone: 1 (614) 292-6679. Email: fredrick.5@osu.edu

Ralf Bundschuh: Phone: 1 (614) 688-3978. Email: bundschuh.2@osu.edu

Supplemental Material A: A test of the shuffling method using randomly-chosen sequences. As a control, for 2000 16S cases and 2000 23S cases, we randomly selected two sequence regions from the genome (hereafter referred to as the mock LT) with the same lengths as the leader and trailer. The shuffling method was applied to these mock LTs, and histograms showing the distribution of z-scores are shown below in **Figure S1**. These histograms clearly show low z-scores, indicating that the high z-scores we observe from real LTs is indicative of LT helices. Furthermore, of the 2000 16S rRNA mock LTs, 90 (4.50%) have a z-score of above 2 and for the 2000 23S rRNA mock LTs, 85 (4.25%) have a z-score of above 2. These numbers correspond closely with the expected p-value of 0.05.

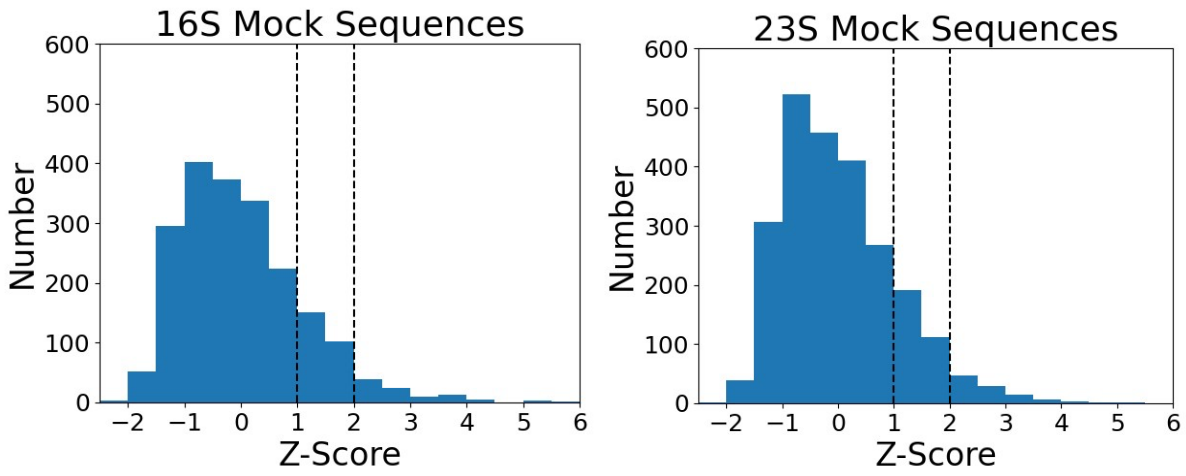


Figure S1. Distribution of mock sequence z-scores for the shuffling method.

Supplemental Material B: Successful sliding window threshold for covariation method. To obtain a logical threshold for the number of sliding windows meeting the LT base pairing criteria for the covariation method, we investigated the distributions of the shuffling method number of SSWs corresponding to a z-score of 1 for each of the LTs in our study in **Figure S2**.

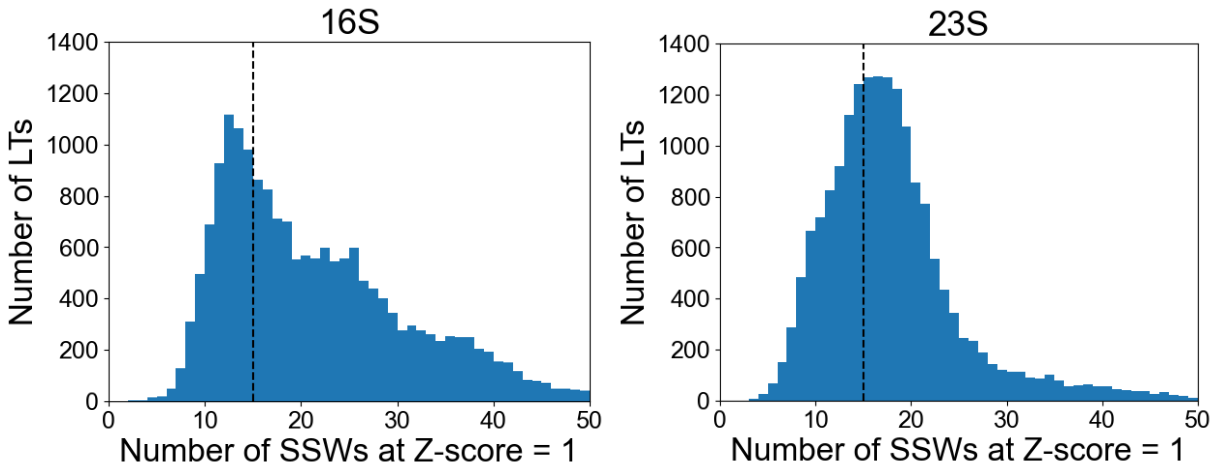


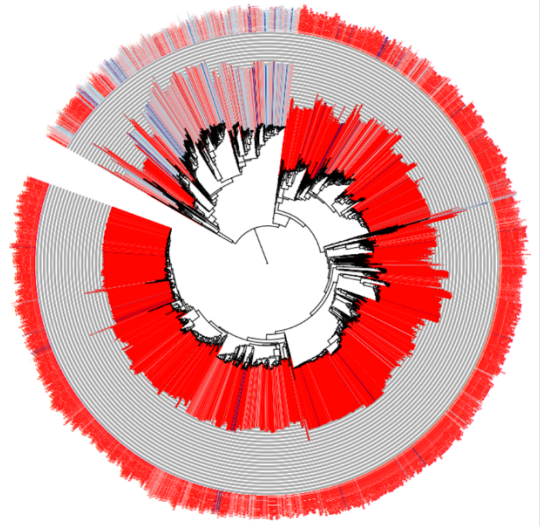
Figure S2. Number of SSWs corresponding to a z-score of 1 in the shuffling method.

Across both 16S and 23S LTs, the number of SSWs at a z-score of 1 peaks at approximately 15. Combined with the number of SSWs histograms for the covariation method for all LTs (**Figure 2C** for 16S and **Figure 4B** for 23S) peaking after a value of 15, we rationalize that 15 is a logical threshold to define as whether or not the LT has a predicted stem. We note that the distributions in **Figure S2** were generated using structures from the shuffling method, while we expect structures generated from the covariation method to contain a lower number of SSWs due to divergent structure regions. This makes the threshold of 15 a somewhat more stringent one for the covariation method than a z-score of 1 for the shuffling method.

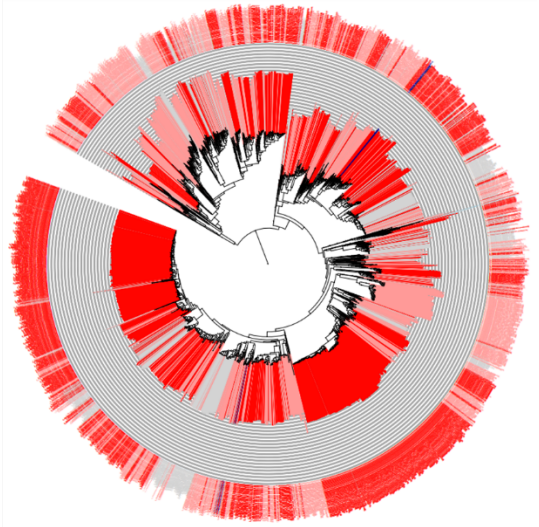
Supplemental Material C: Phylogenetic trees demonstrate the prevalence of 16S and 23S LT helices for each phylum using the default shuffling and covariation methods. Phylogenetic trees of all phyla shown in Figures 3 and 5 for 16S and 23S respectively are shown below in **Figure S3**. Each LT leaf node is colored by shuffling-method z-scores (dark blue: no signal, light blue: weak signal, light red: intermediate signal, dark red: strong signal) or covariation-method results (dark blue: no evidence for stem with strong covariation, light blue: no evidence for stem with weak covariation, light red: evidence for stem with weak covariation, dark red: evidence for stem with strong covariation). Grey leaf nodes signify no prediction. High resolution trees are available as PDFs on 10.5281/zenodo.11051088.

Proteobacteria 16S

Shuffling Method

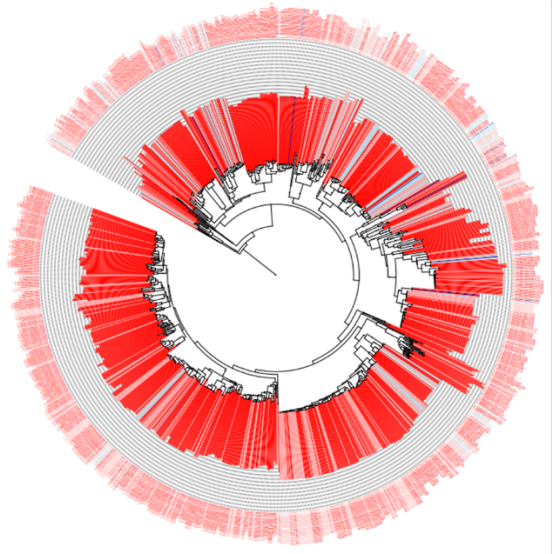


Covariation Method

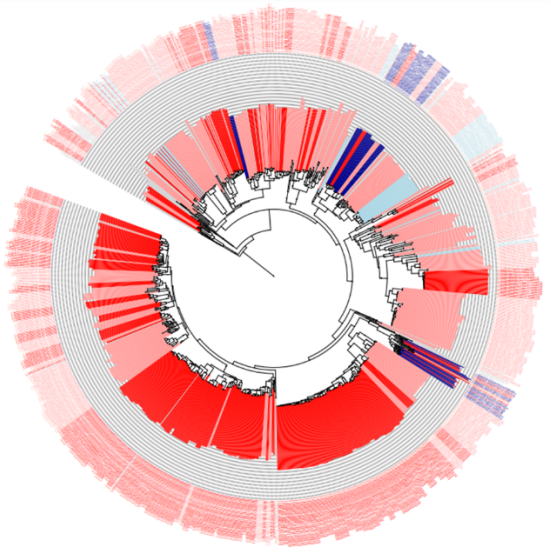


Bacteroidota 16S

Shuffling Method

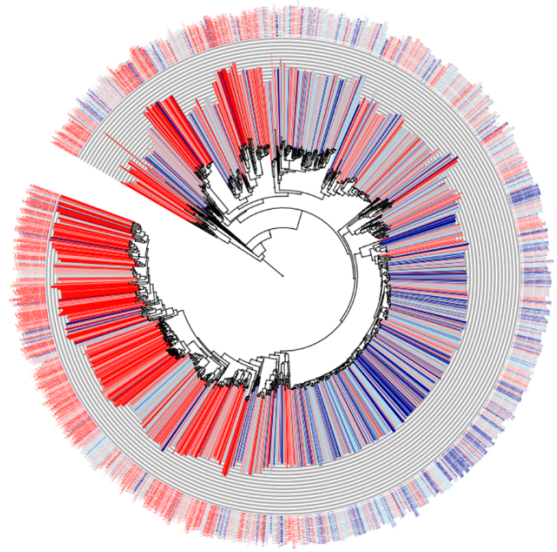


Covariation Method

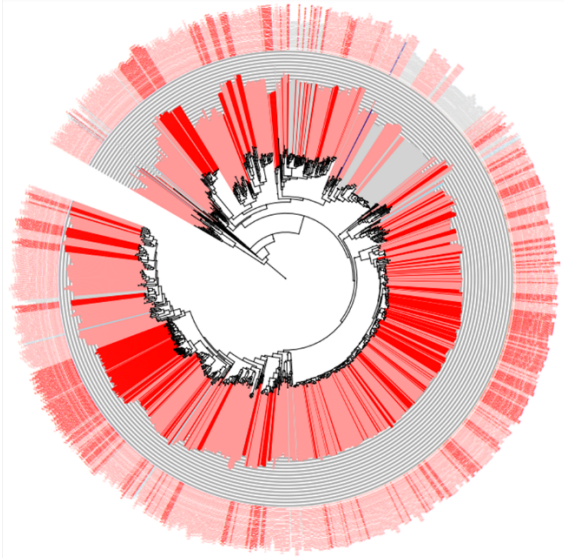


Actinobacteriota 16S

Shuffling Method

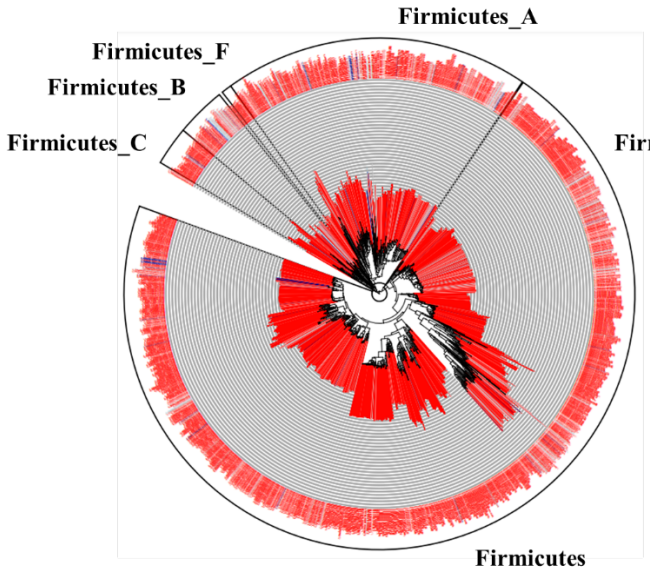


Covariation Method

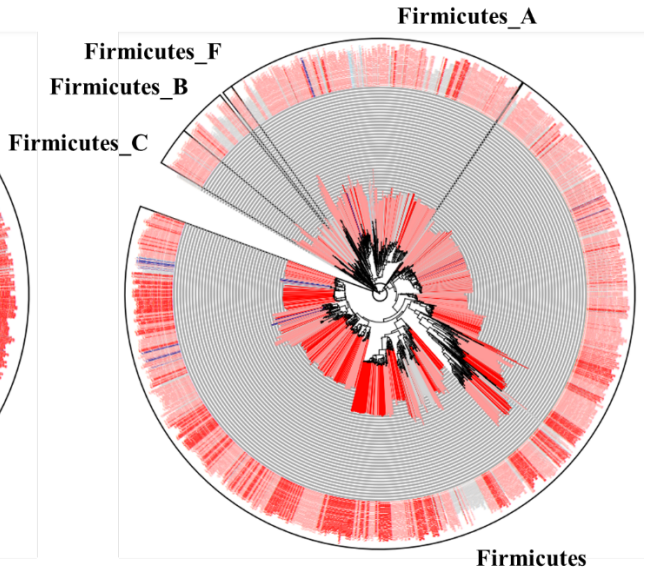


Firmicutes 16S

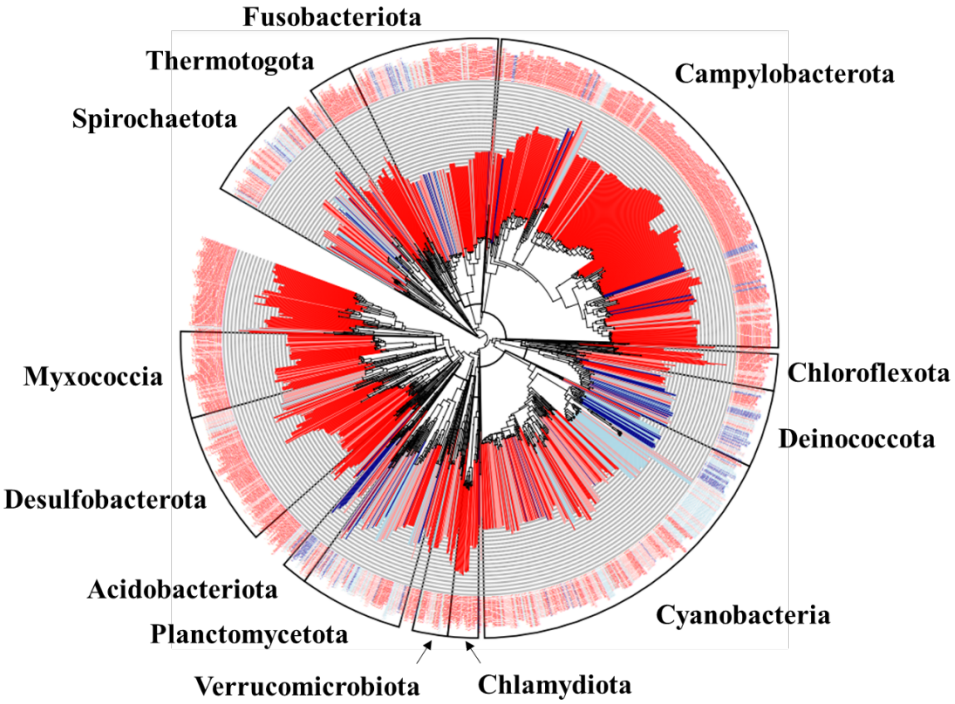
Shuffling Method



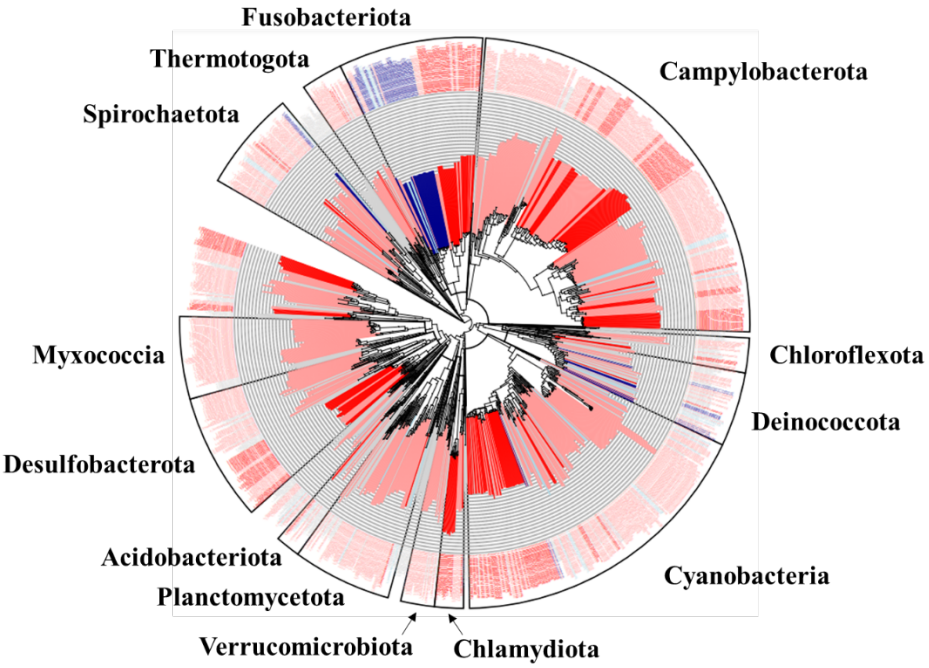
Covariation Method



Remaining Bacterial Phyla 16S – Shuffling Method

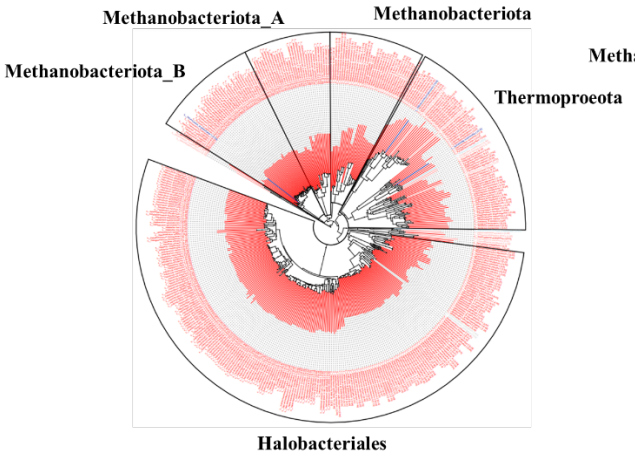


Remaining Bacterial Phylum 16S – Covariation Method

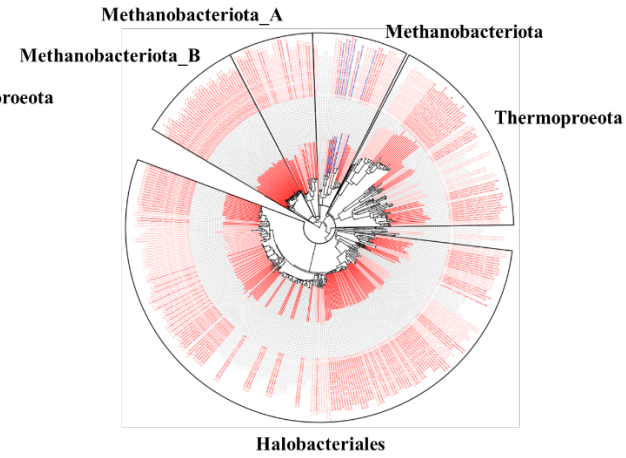


Archaea 16S

Shuffling Method

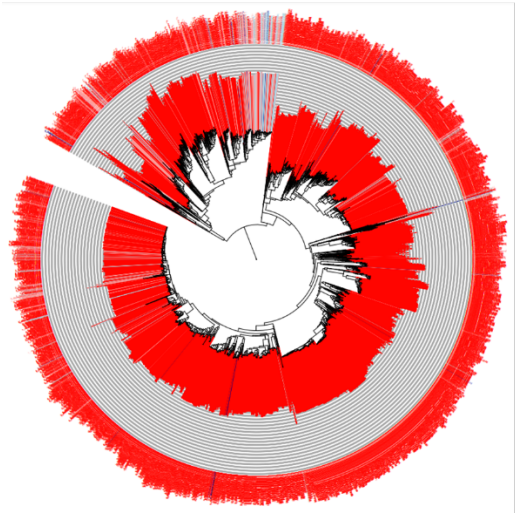


Covariation Method

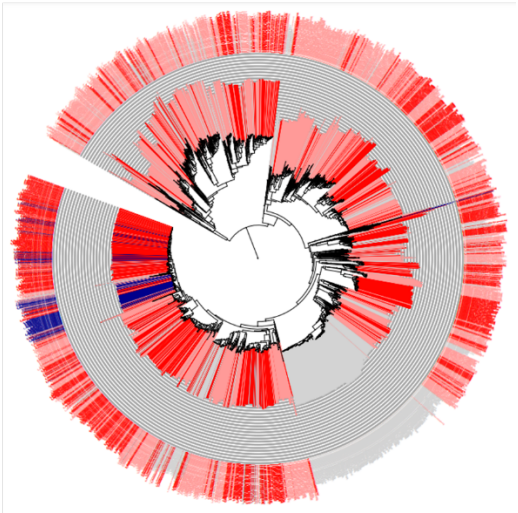


Proteobacteria 23S

Shuffling Method

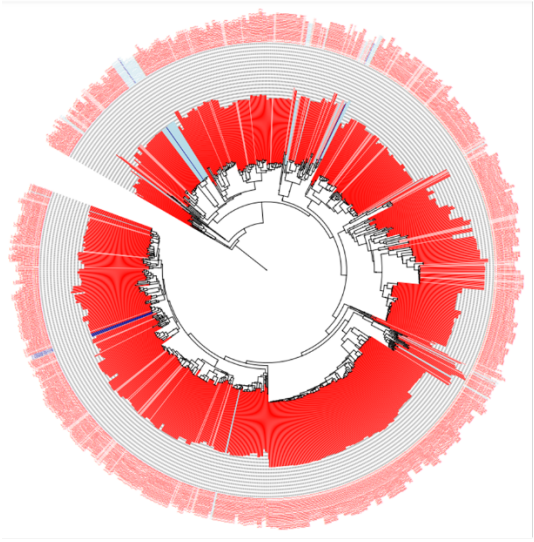


Covariation Method

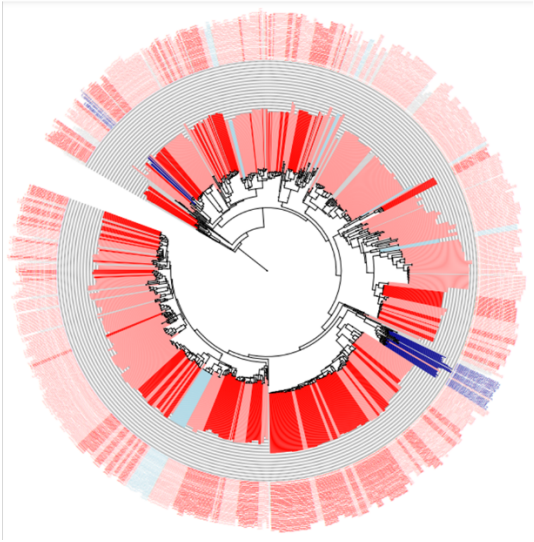


Bacteroidota 23S

Shuffling Method

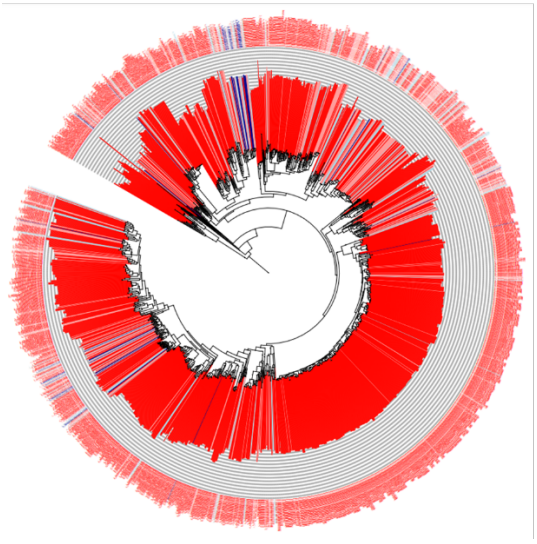


Covariation Method

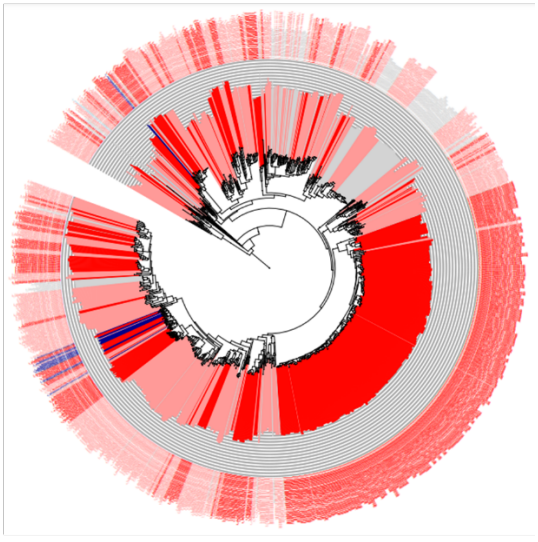


Actinobacteriota 23S

Shuffling Method

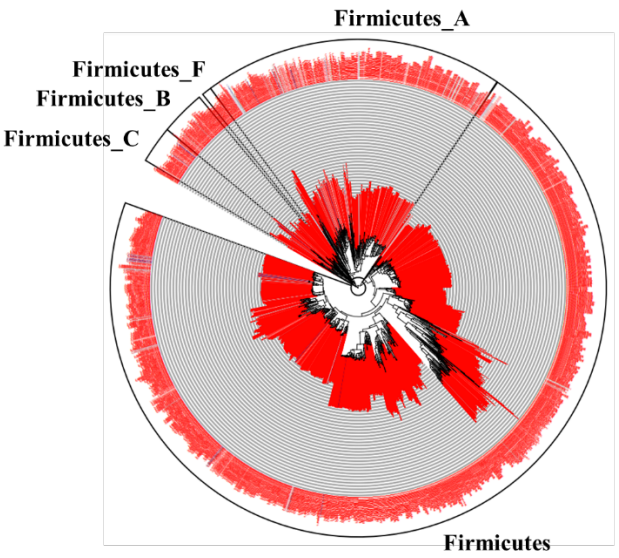


Covariation Method

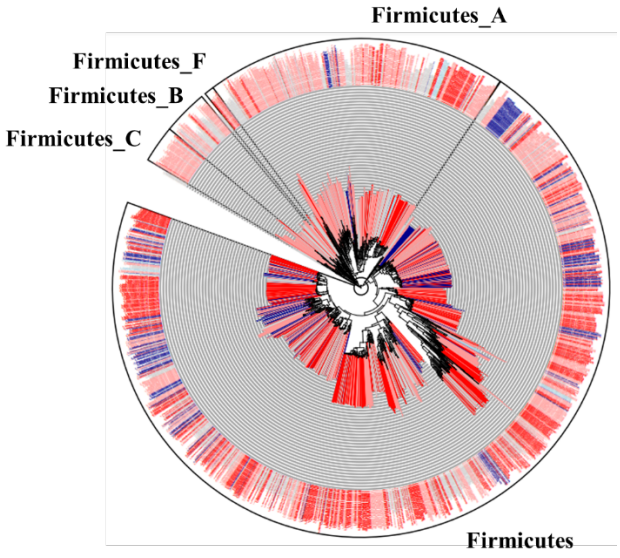


Firmicutes 23S

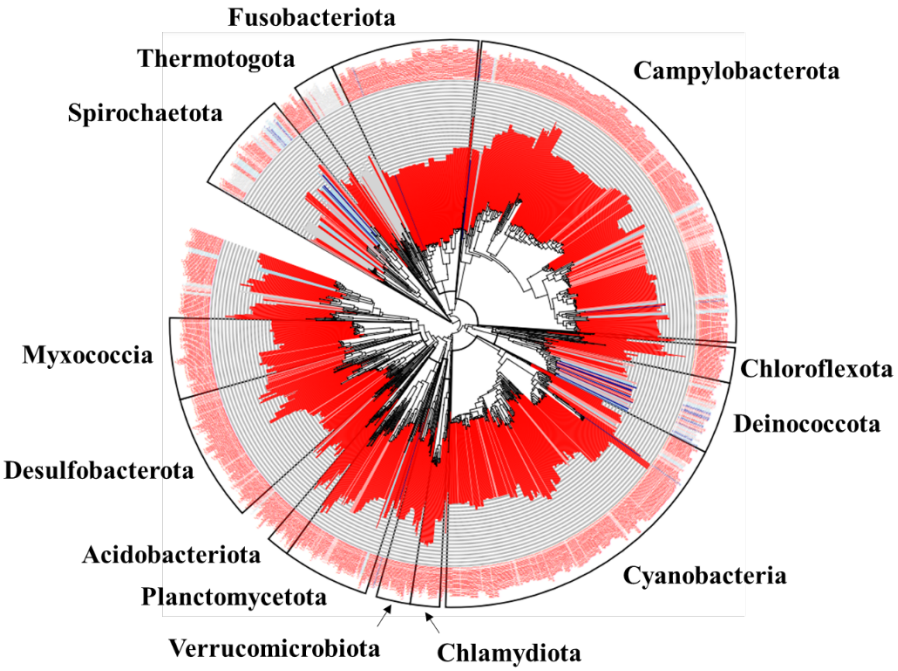
Shuffling Method



Covariation Method



Remaining Bacterial Phyla 23S – Shuffling Method



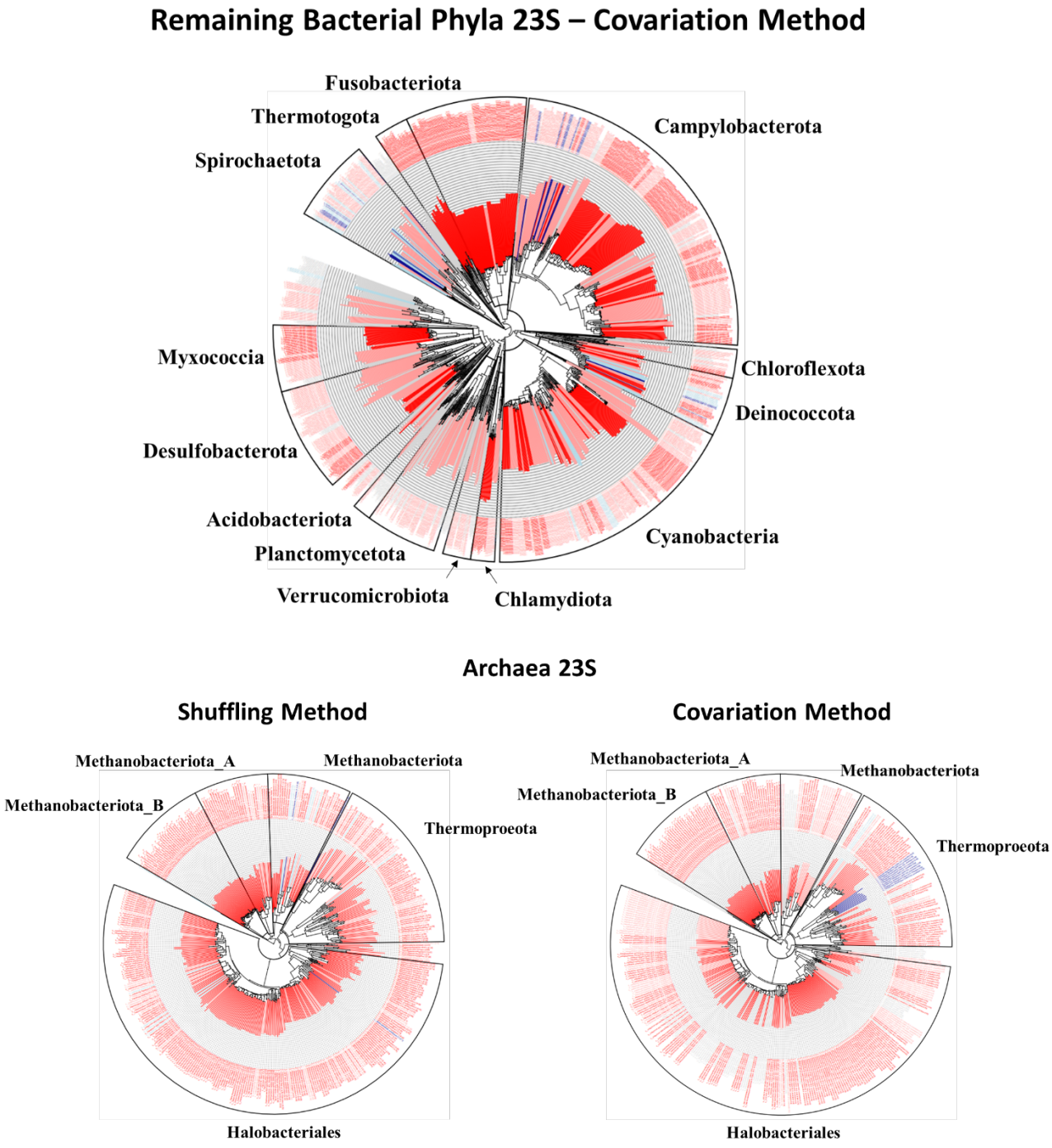


Figure S3. Predicted LT stems in bacterial and archaeal phyla with both the shuffling and covariation methods.

Supplemental Material D: Enlarged versions of structures in figures 6 and 7. Full-page figures of the individual structures shown in **Figure 6** and **Figure 7** are shown below.

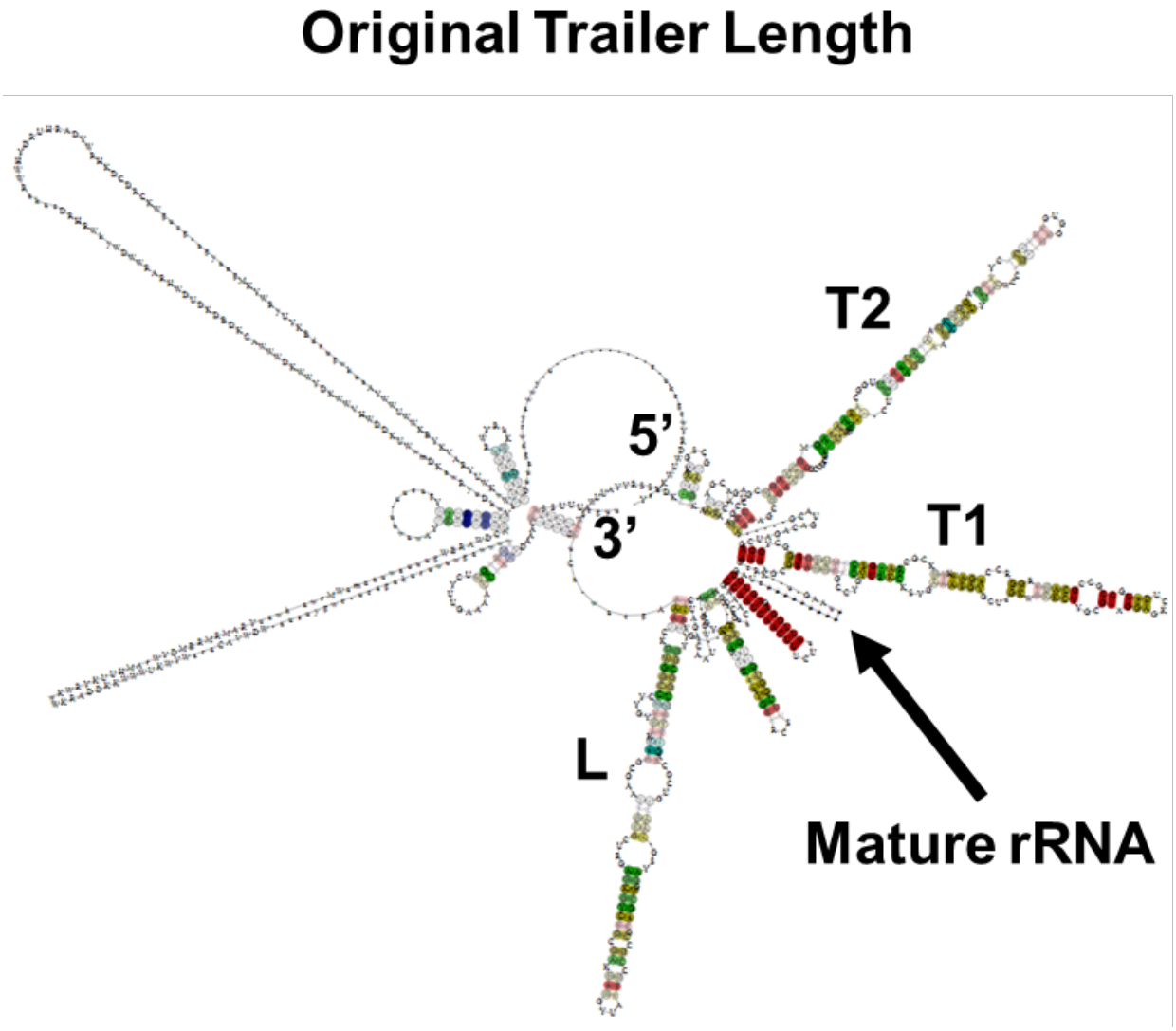


Figure S4. Enlarged structure in Figure 6B.

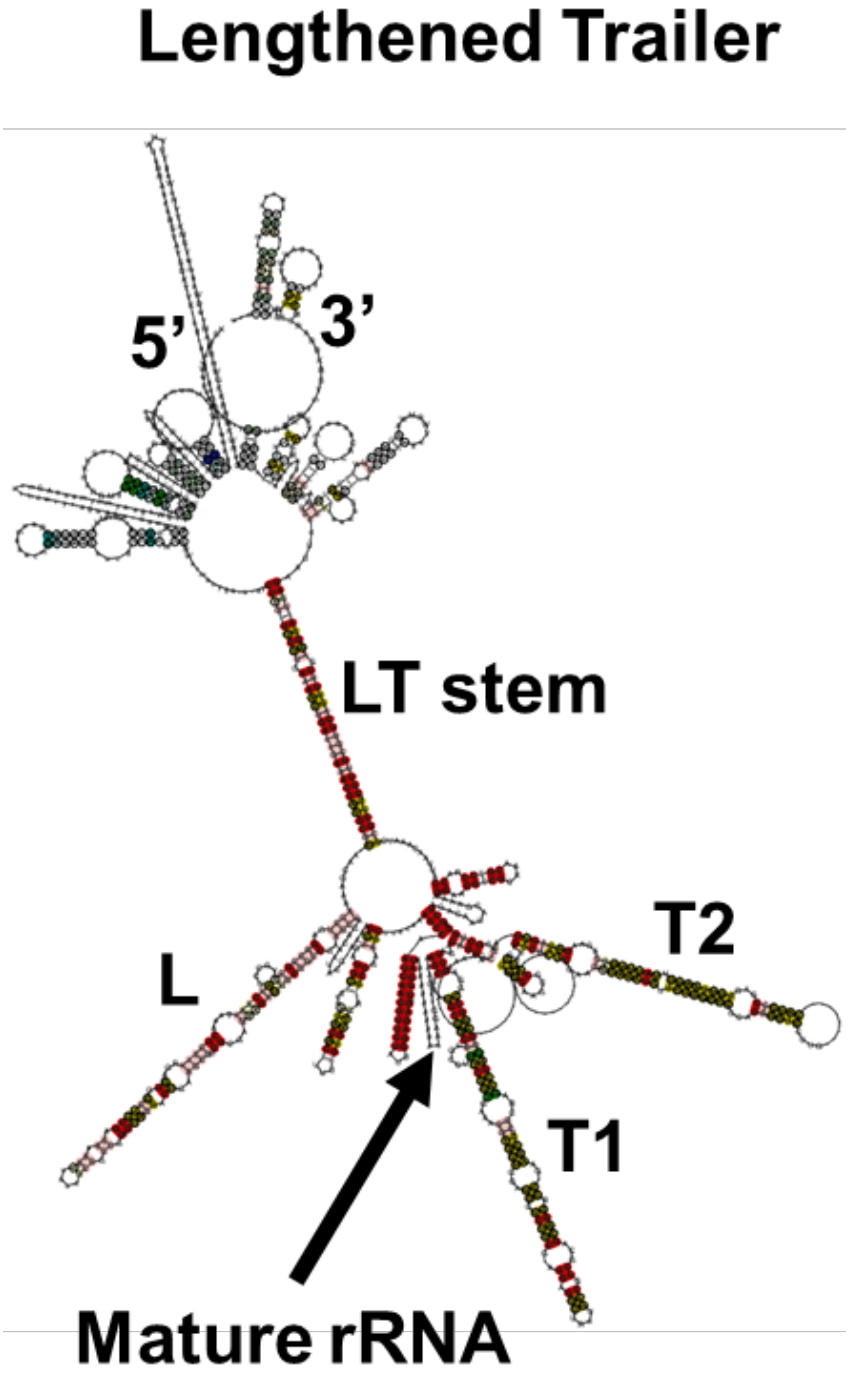


Figure S5. Enlarged structure in Figure 6C.

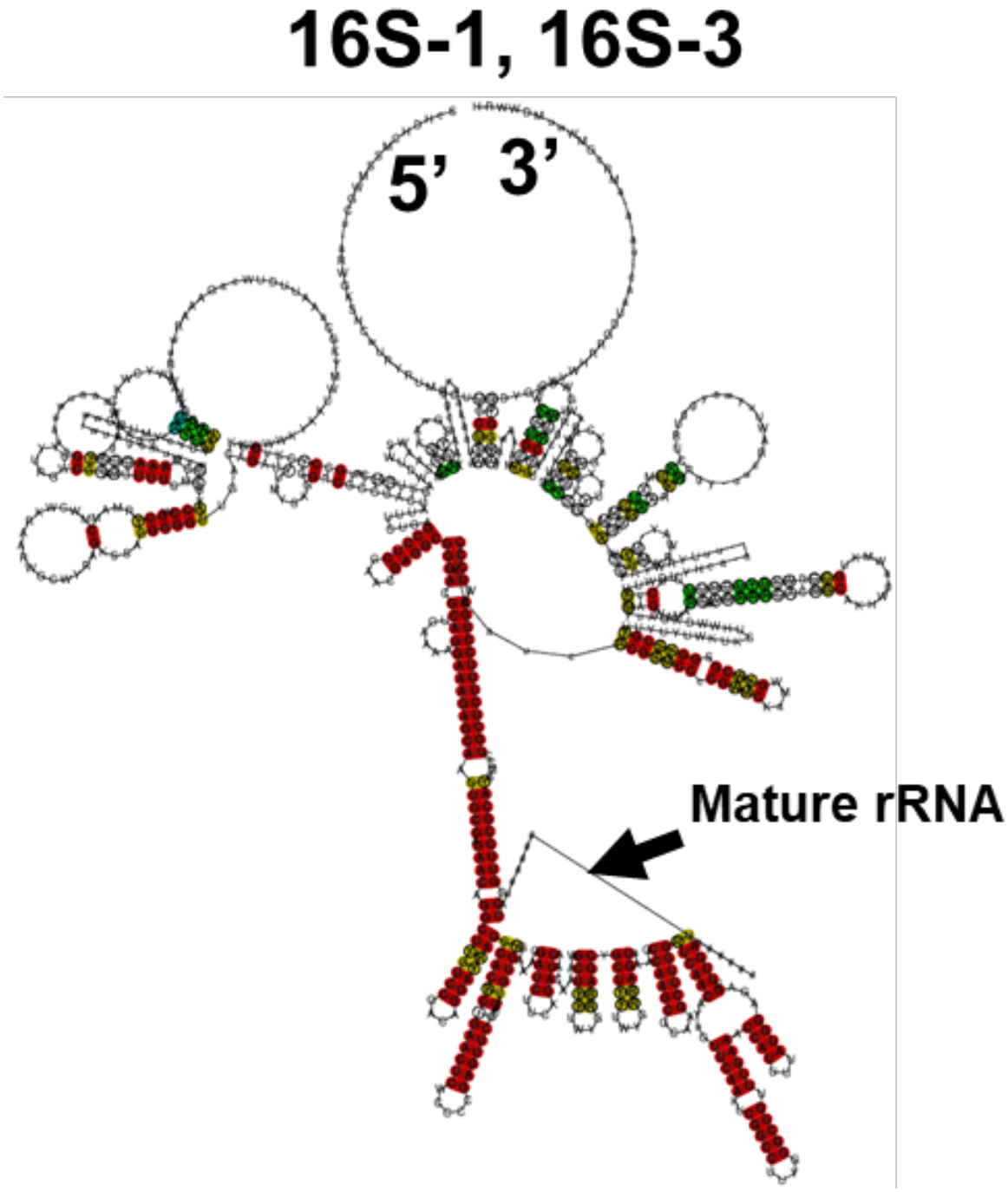


Figure S6. Enlarged structure in Figure 7C, 16S-1 and 16S-3.

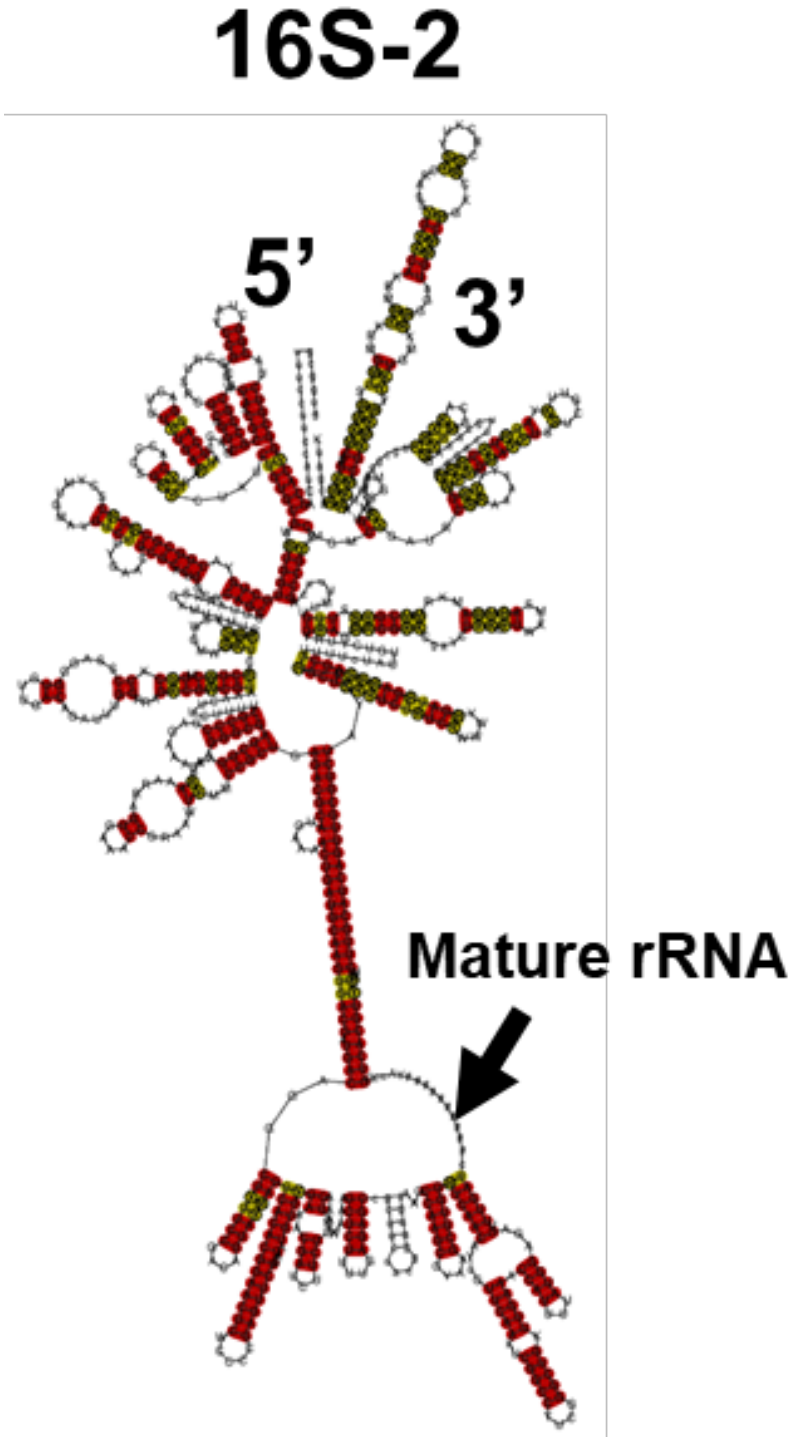


Figure S7. Enlarged structure in Figure 7C, 16S-2.

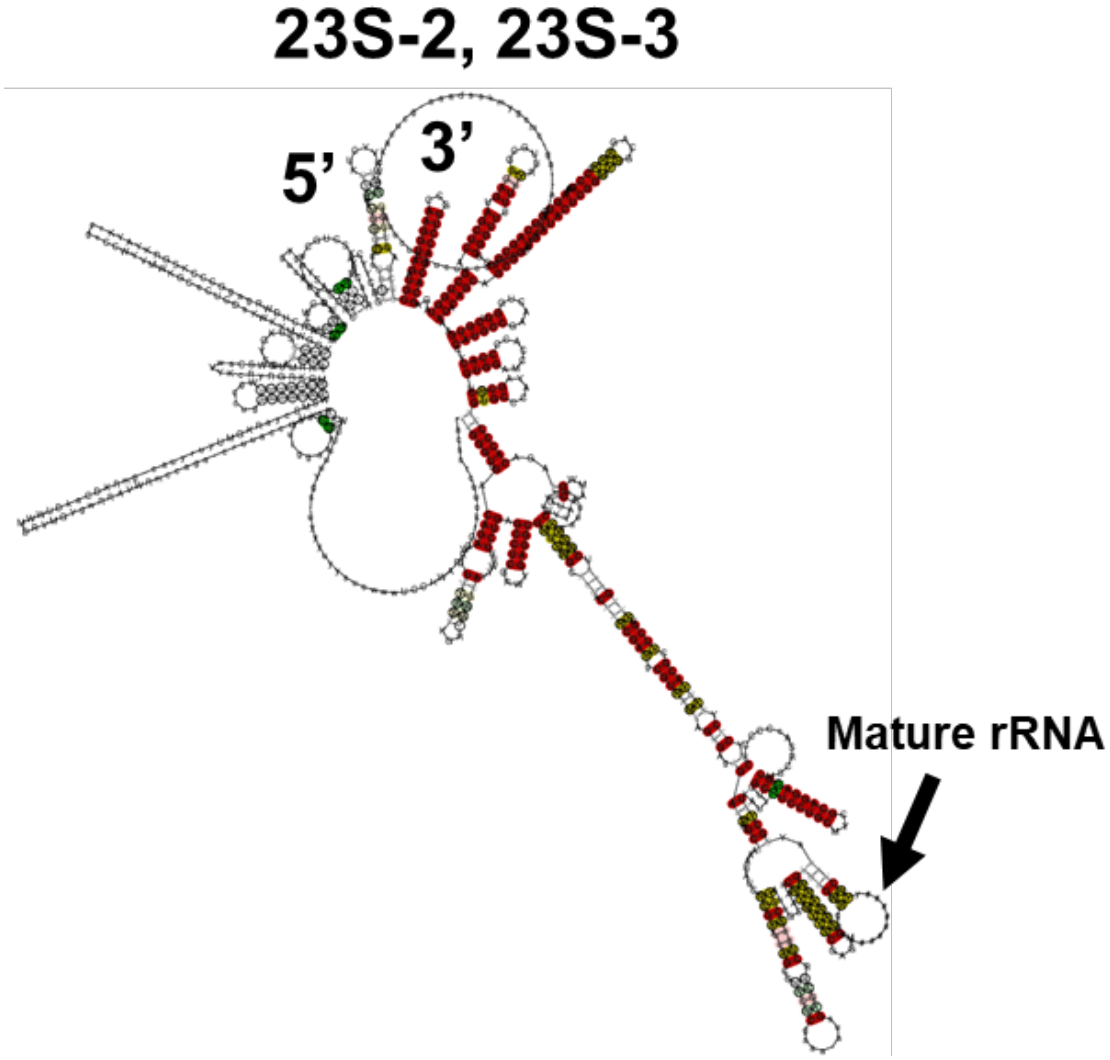


Figure S8. Enlarged structure in Figure 7D.

Supplemental Material E: Investigation of 16S and 23S LTs without a predicted stem. LTs without a predicted stem in either method (corresponding to the phyla whose y-axis fraction was less than 1.0 in **Figure 3(C)** for 16S and **Figure 5(C)** for 23S) were compiled for additional analysis. 443 such 16S LTs and 166 such 23S LTs were identified.

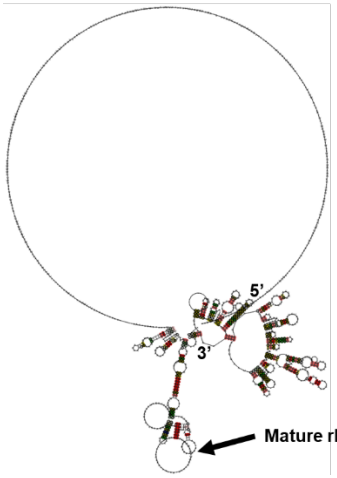
The shuffling method was rerun with the following changes: (1) upstream and downstream gene annotations were ignored when obtaining leader and trailer nucleotides and (2) the trailer length was increased to 500 nucleotides downstream of the mature rRNA annotation. Under these relaxed conditions, 101 of 443 16S LTs and 47 of 166 23S LTs contained a z-score over 1.0, corresponding to a predicted LT stem.

The remaining LTs not containing a predicted LT stem (342 16S and LTs and 119 23S LTs) were reanalyzed using the covariation method with the following changes: (1) the trailer length was extended to 500 nucleotides downstream of the mature rRNA annotation and (2) all LT sequences were collectively clustered by sequence similarity (independently for 16S and for 23S), ignoring taxonomic class. Under these conditions, 42 of 342 16S LTs and 49 of 119 23S LTs contained a number of successful sliding windows (SSWs) above 15 in their consensus structures, corresponding to a predicted LT stem.

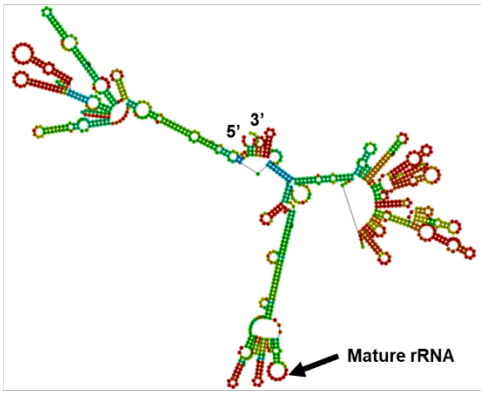
The remaining 300 16S LTs and 70 23S LTs were randomly subsampled for individual manual analysis to determine if their LTs contained no stem or if something about them caused the methods to produce false negatives. Using the extended trailer described above, the consensus structures from the covariation method and the RNAFold minimum free energy structure were compiled. **Figure S9** shows 3 randomly selected 16S LTs and 3 randomly selected 23S LTs, where the position of the mature rRNA is marked by an arrow.

Leptotrichia wadei 16S #1

Covariation Structure

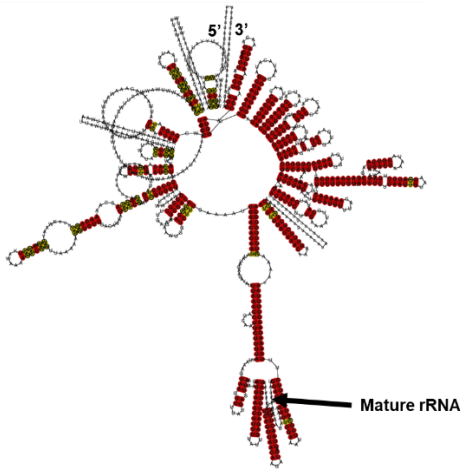


RNAFold Structure

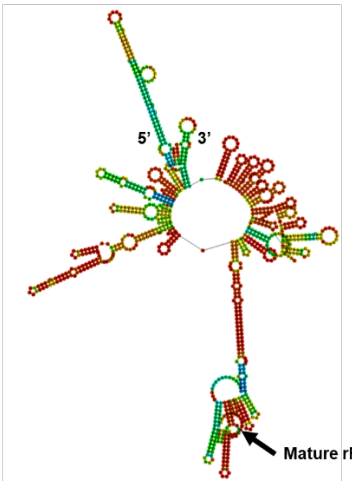


Clostridium baratii 16S #3

Covariation Structure

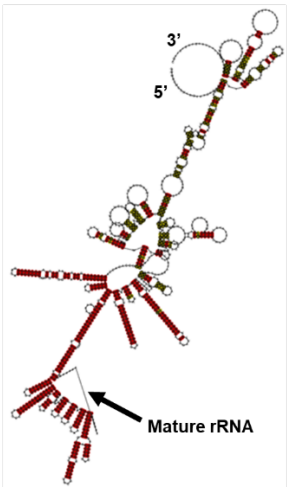


RNAFold Structure

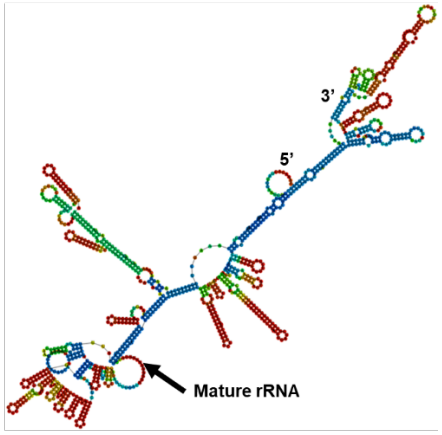


Deinococcus wulumuqiensis 16S #1

Covariation Structure

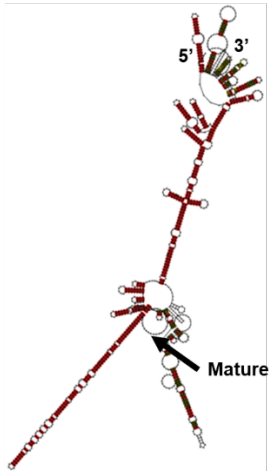


RNAFold Structure

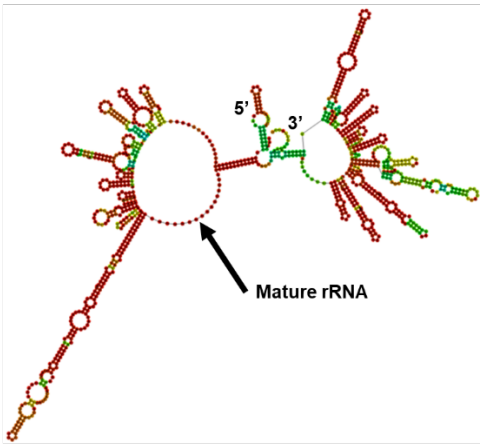


Selenomonas sputigena 23S #1

Covariation Structure



RNAFold Structure



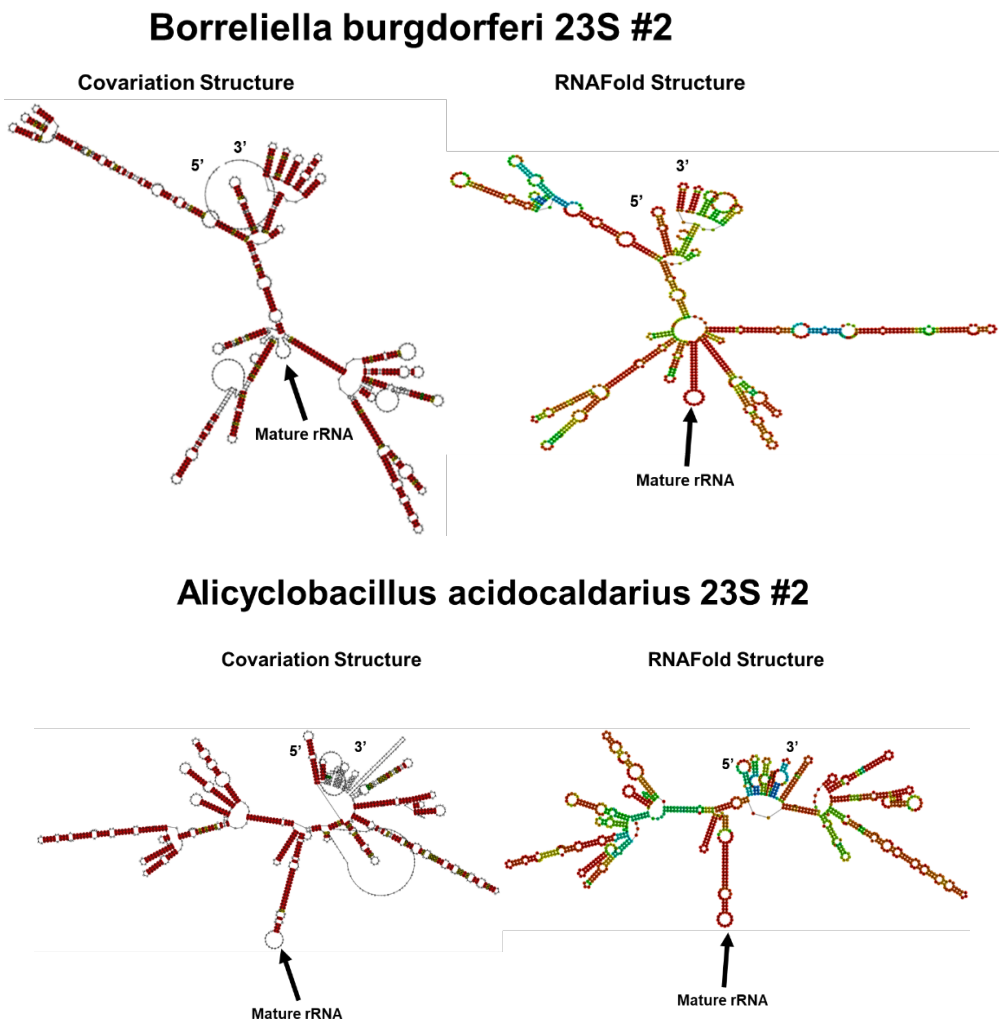


Figure S9. Individual analysis of 16S and 23S LTs with extended trailers reveals LT stems.

It is evident from visually examining all the structures that a LT stem is present. In the majority of cases, the LT stem is interrupted by internal loops and/or bulges, which disrupt the sliding window based method we implemented to quantitatively represent the presence of a LT stem by preventing consecutive runs of 10 nt sliding windows in both the leader and trailer from containing 8 LT base pairs. Thus, we argue that these LTs without a predicted stem are very likely false negatives in both methods and therefore conclude that LT stems are virtually ubiquitous across both 16S and 23S, even in species previously thought not to contain a LT stem (such as *Deinococcus radiodurans*).

Supplemental Material F: Additional evidence for ubiquity of 16S and 23S LT helices. For each species, all 16S LTs and 23S LTs were analyzed to determine what fraction of LTs in a given species contain a predicted LT stem in either method (default method data shown, corresponding to **Figure 3** for 16S and **Figure 5** for 23S). Histograms of these fractions are shown in **Figure S10**.

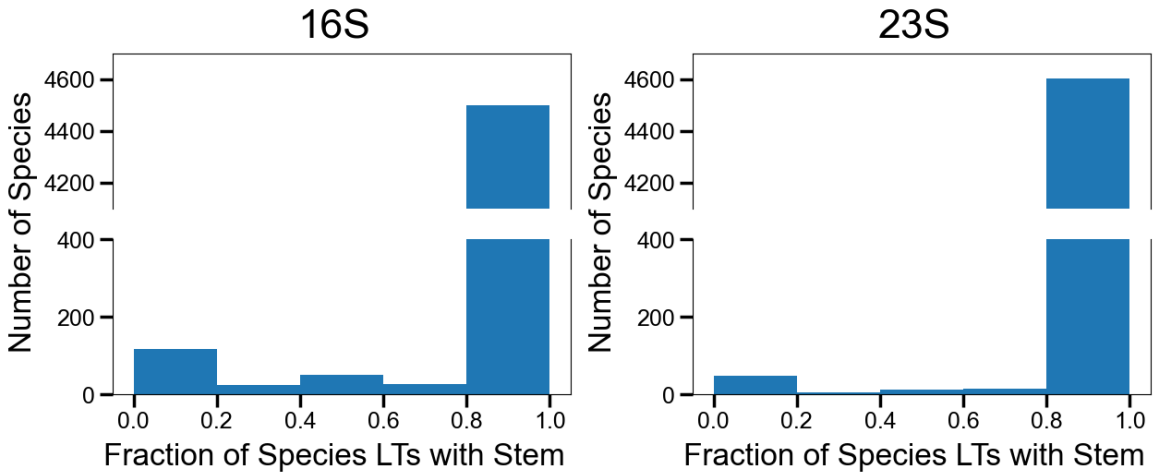


Figure S10. Fraction of LTs with a predicted stem in each species.

A small subset of species (126 16S and 47 23S) contain multiple LTs, some with a predicted stem and some without. For these species, LT stems are retained in at least some rRNA genes, consistent with a critical function role. Of the species for which none of the LTs contained a stem prediction (16S: 115 species with 214 LTs, 23S: 47 species with 87 LTs), the number of LTs per species was normalized to sum to 1 and compared to the normalized number of LTs per species for all species. The normalized value for species with no LT stem predictions was divided by the normalized value for all species to quantify the enrichment of LT numbers in species without any predicted LTs over the background. These are shown in **Figure S11** with a black dotted line at y=1.0 signifying an equal proportion in both data sets.

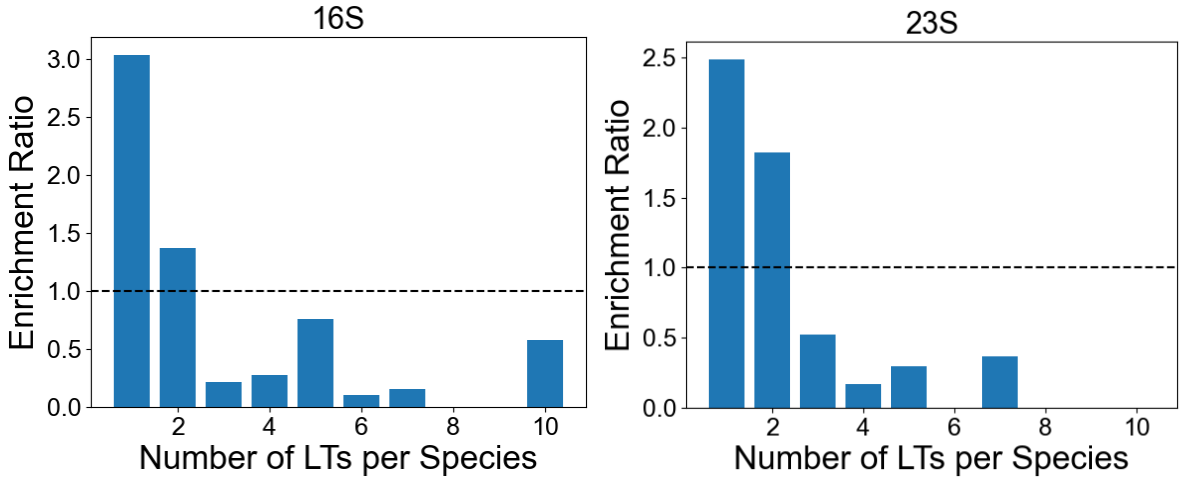
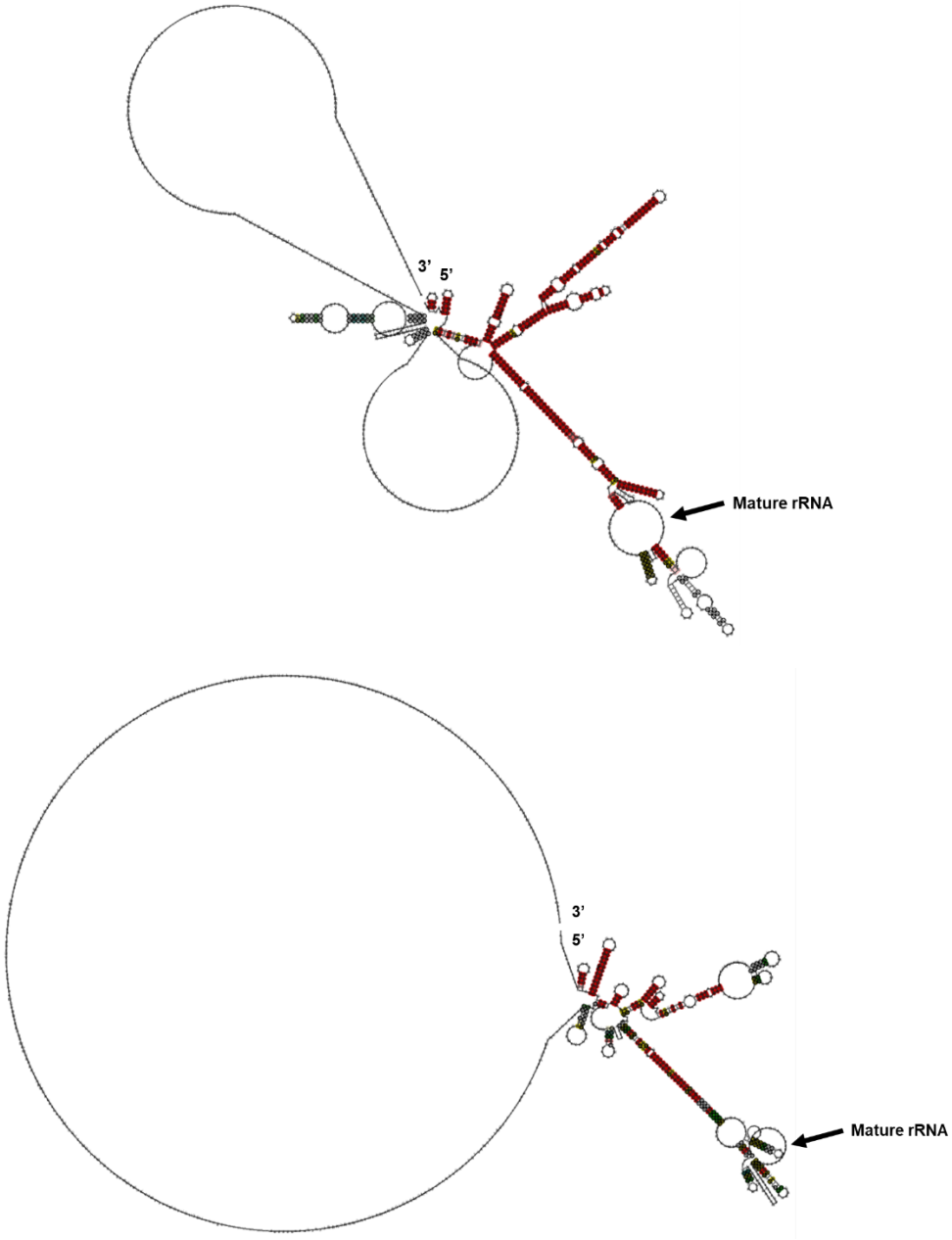


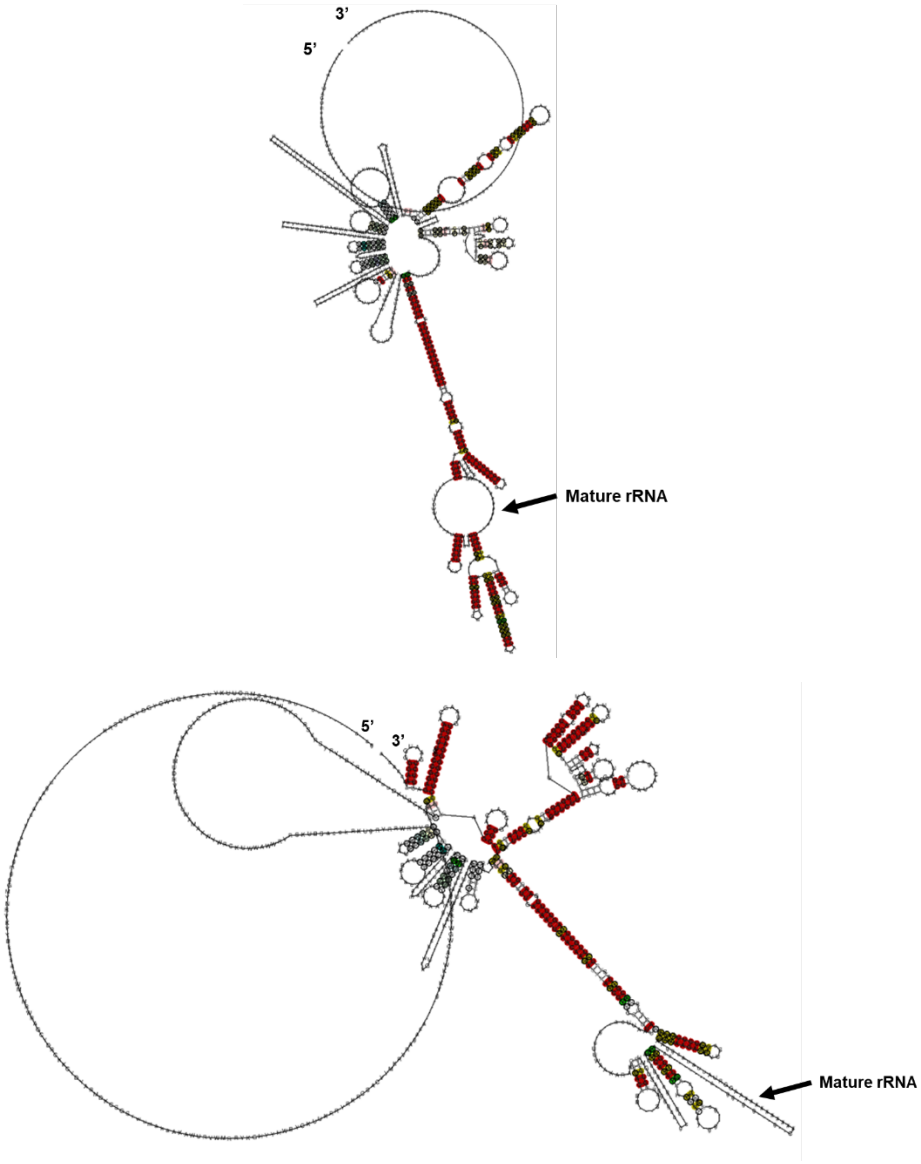
Figure S11. Number of LTs per species enrichment of species with no predicted LT stems versus all species.

The majority of species without any predicted LT stems contain only 1 or 2 16S or 23S rRNA annotations, a difference from the number of LTs per species across all bacteria and archaea. We argue that it is likely by chance that these species do not contain any LT with a predicted stem given the low numbers of LTs per species, and thus they are false negatives.

If we require that a species have no LT stem prediction across any of its 16S and 23S LTs, there are only 10 species containing 24 total LTs (across both 16S and 23S rRNA). We argue that this low number of species (10 of 4,724) are likely false negatives. Alternatively, we hypothesize that these species could have alternative analogous mechanisms to bridge the upstream and downstream regions of pre-rRNA such as a protein bridge.

Supplemental Material G: Species contains high numbers of LT structures across operons. The unique LT clusters from the original covariation method associated with *Bacillus subtilis* were compiled, resulting in 5 unique LT clusters from its 10 16S rRNAs. The 5 unique 16S LT structures are shown in **Figure S12**, where the position of the mature rRNA is marked by an arrow.




3’
5’
Mature rRNA
5’
3’
Mature rRNA

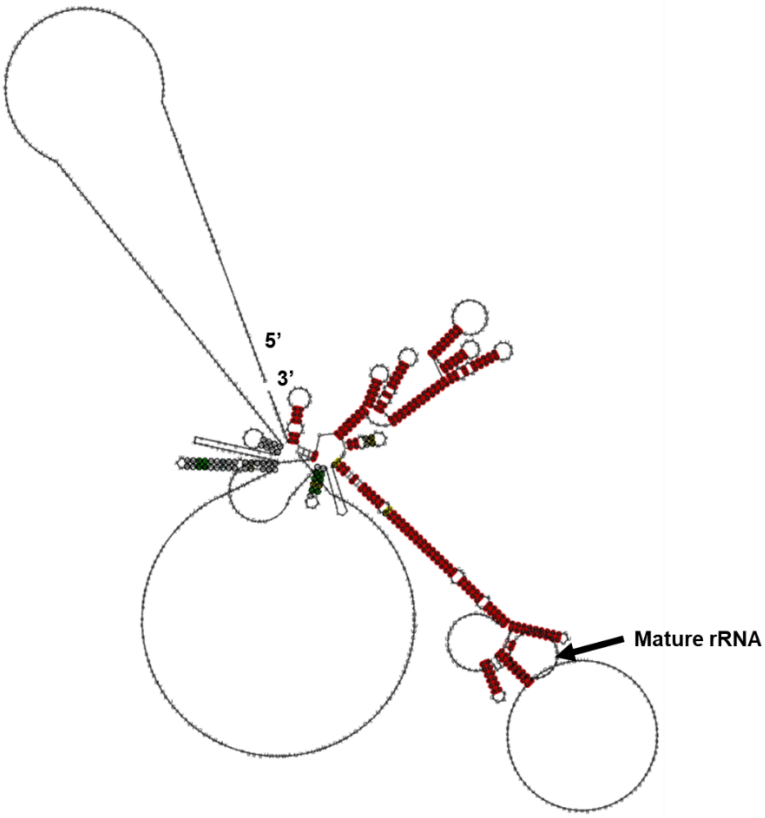


Figure S12. *Bacillus subtilis* 5 unique 16S LT structures from the covariation method.

Despite each of the 5 LT structures looking quite different from one another, the same strong LT helix remains in all structures. We note that structural differences are observed both far away from and close to the mature rRNA.

Curiously, while *Bacillus subtilis* contains 5 unique LT structures for its 10 16S rRNAs, the species contains only 2 unique LT structures for its 10 23S rRNAs (**Figure S13**).

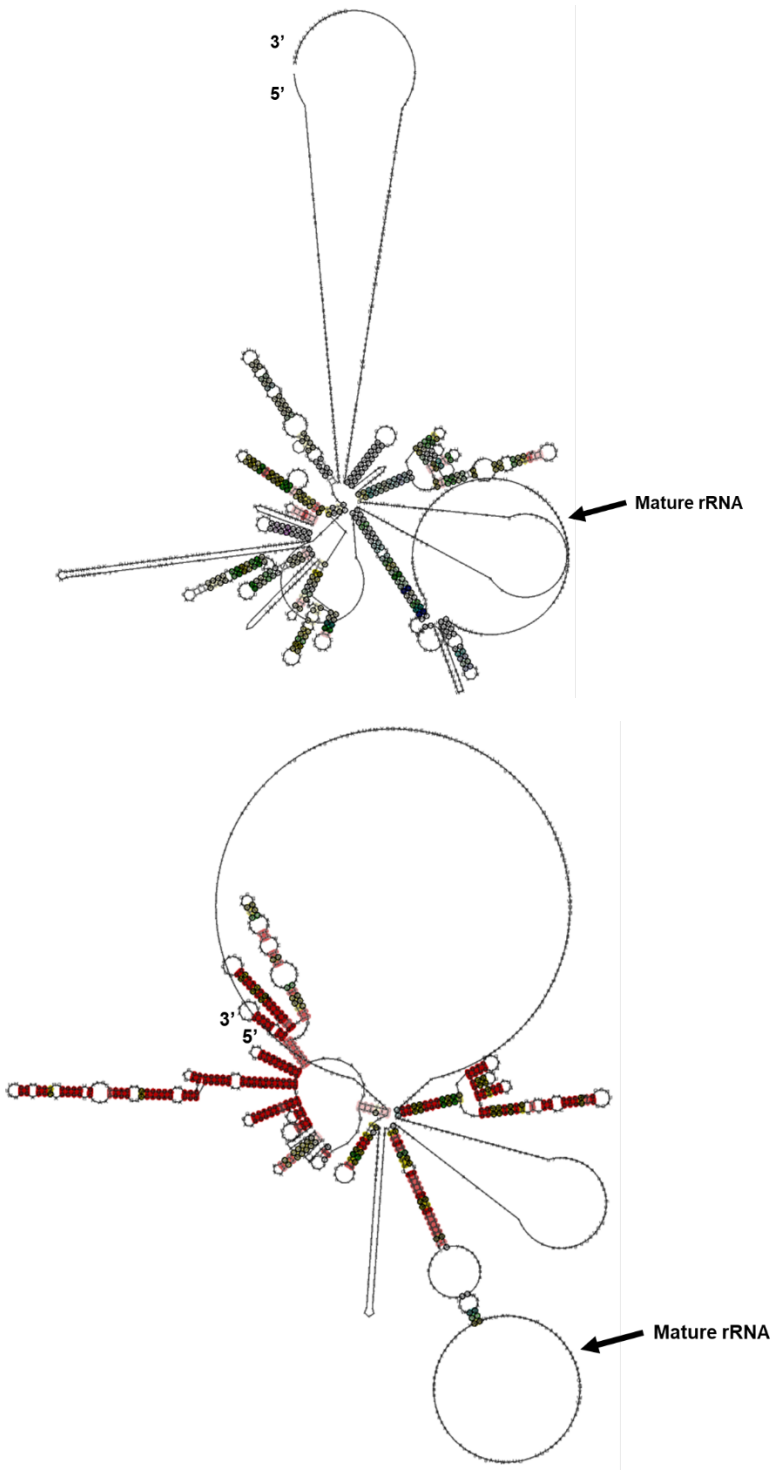


Figure S13. *Bacillus subtilis* 2 unique 23S LT structures from the covariation method.

Supplemental Material H: Species LT structure diversity. For each species, the total number of unique LT clusters from the original covariation method associated with the species were compiled. We note that the number of unique LT clusters included clusters of size 1 (e.g., those for which RNAClust was not run subsequently); however, rRNAs without enough up- and down- stream nucleotides for the covariation method under default settings were not counted in the analysis. The total number of LTs per species was compared to the number of unique LT clusters per species to determine how diverse the LTs are for each species in **Figure S14**.

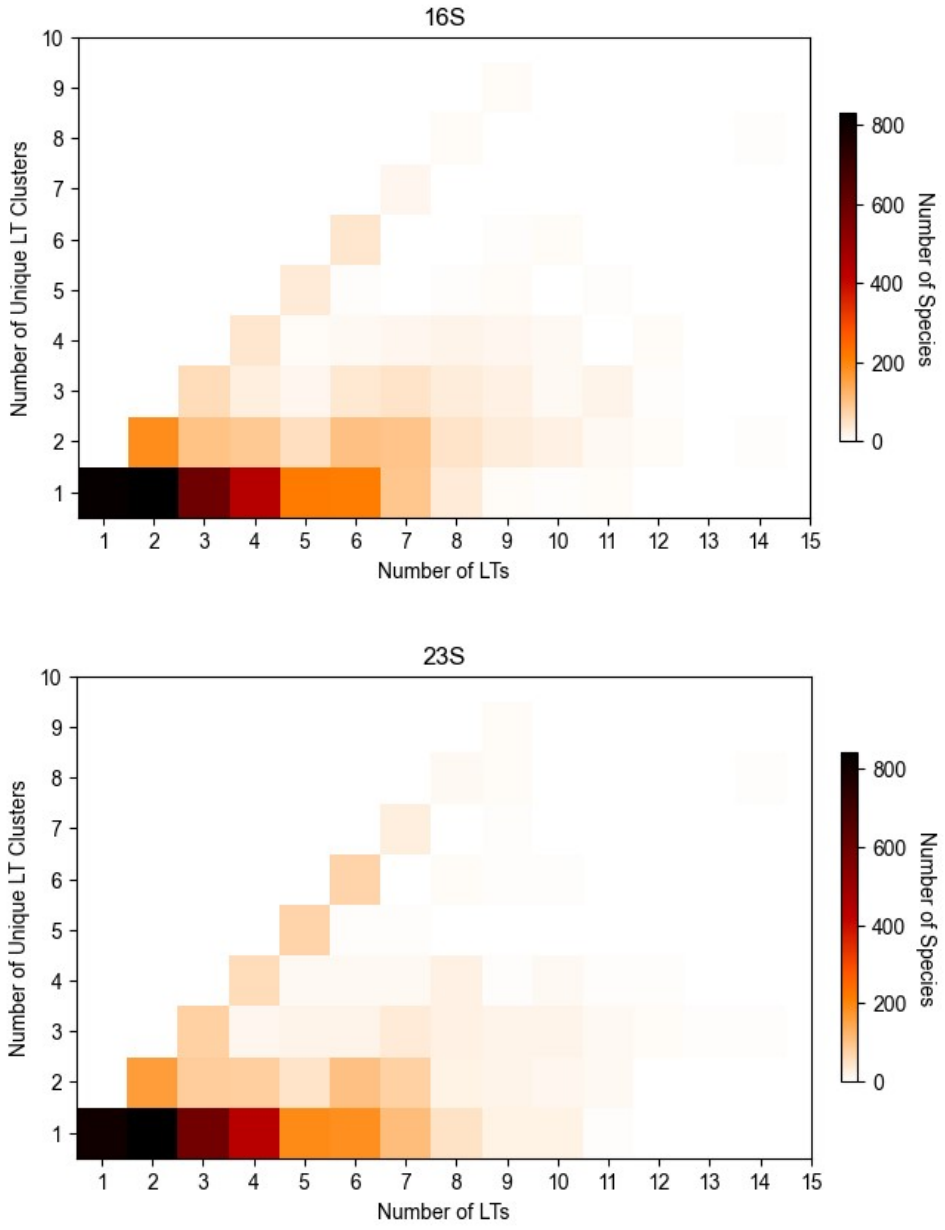
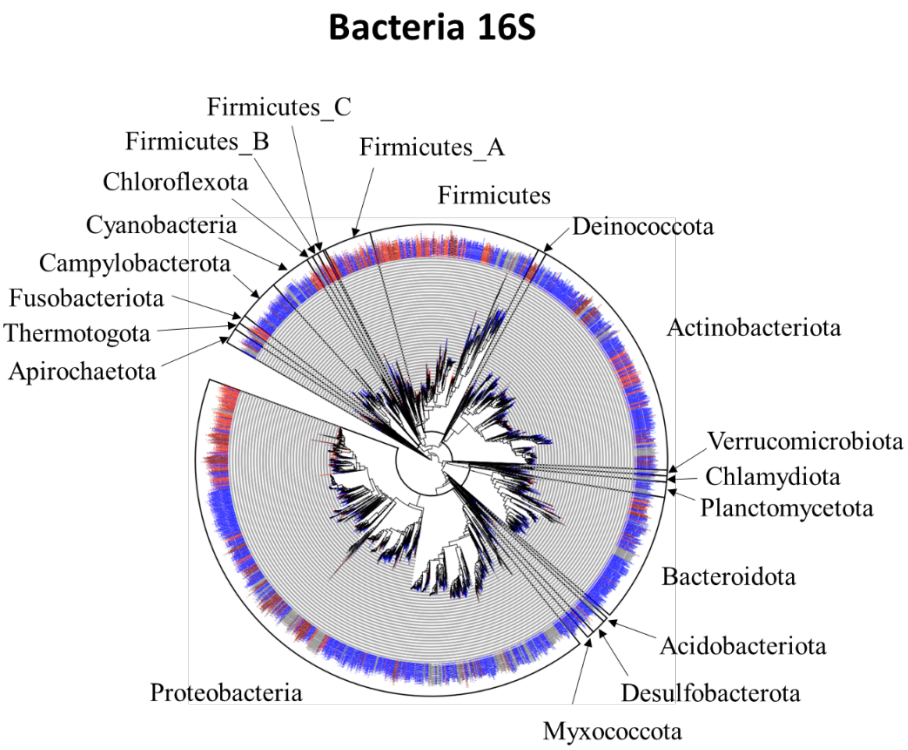


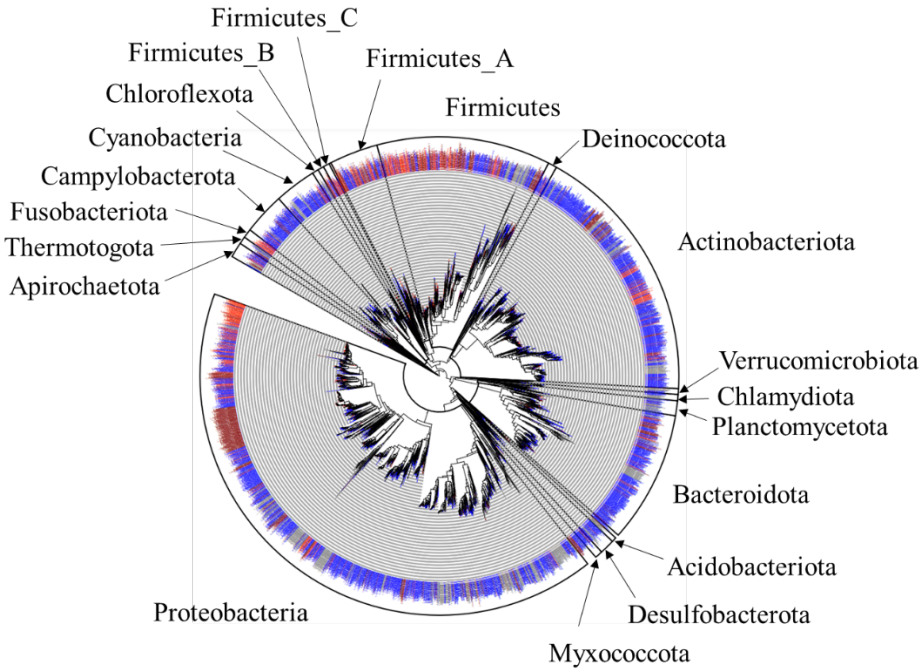
Figure S14. Number of LTs versus number of LT clusters. The heatmaps were generated with MatPlotLib's "clf" function. 0.35% and 0.44% of 16S and 23S LTs, respectively, fall outside of the plot windows shown

The majority of species contain a relatively low number of both 16S and 23S LTs and only 1 LT cluster. A subset of species contains more than 1 LT cluster (1,361 16S species and 1,322 23S species), signifying that variability of LTs within the same species is fairly common. Interestingly, in addition to the many species with one or two LT clusters, we observe a prominent diagonal in the heatmap, i.e., many species for which each LT is in a different cluster even if there are up to nine LTs in a species.

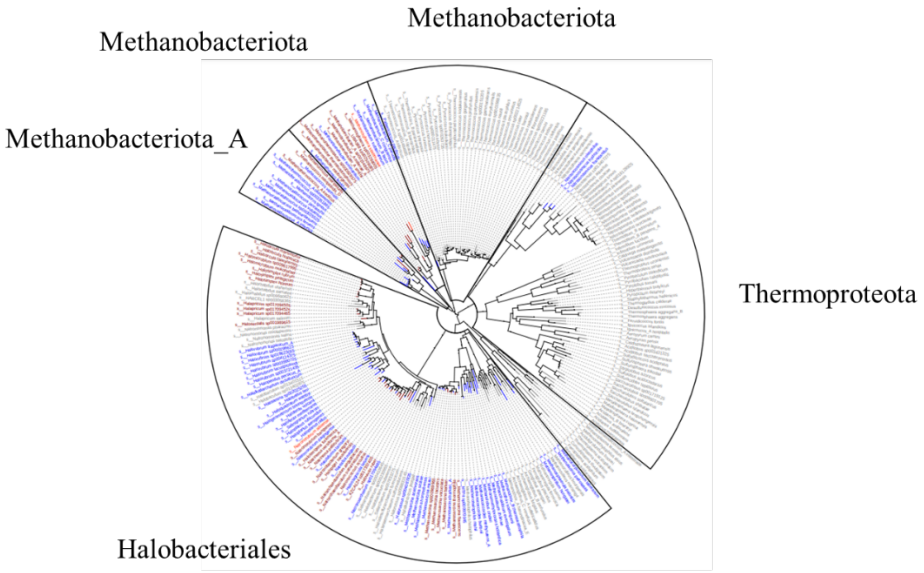
To find out if diversity of LTs within individual species is limited to specific lineages, bacterial and archaeal phylogenetic trees (**Figure S15**) were created that quantify what fraction of a species' LTs belonged to unique clusters. Leaf nodes represent species; they are colored grey if the species contains only 1 LT, else blue if the species contains only 1 unique LT clusters or the fraction is ≤0.25, else light red if the species fraction is ≤0.50, else maroon if the species fraction is >0.50. For bacteria, we find significant concentration of species with multiple LT clusters for Enterobacterales and Bacilli in both 16S and 23S.



Bacteria 23S



Archaea 16S



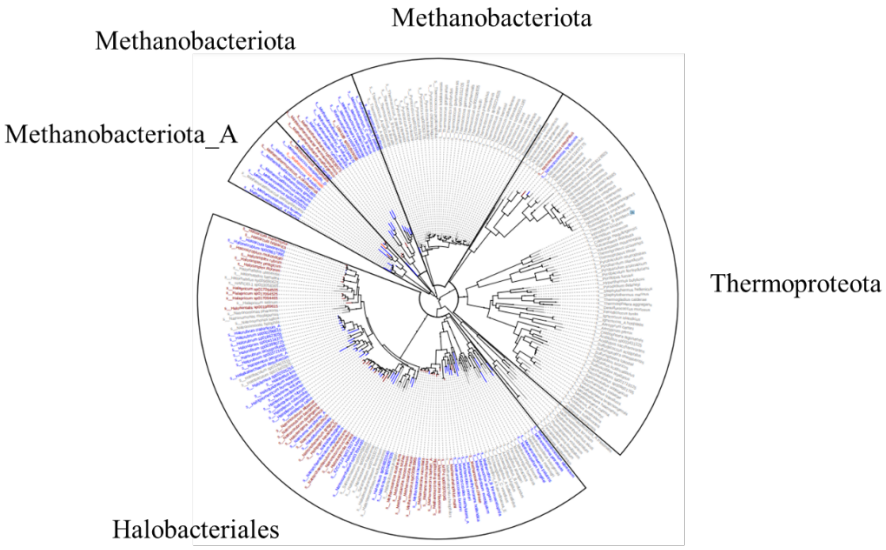


Figure S15. Phylogenetic distribution of species with many LT clusters.