

Analysis of 1,000 Type-Strain Genomes Improves Taxonomic Classification of Alphaproteobacteria

Supplementary File S2

List of Figures

1	Complete phylogenomic GBDP tree of the <i>Alphaproteobacteria</i> genome dataset	16
2	Backbone-constrained 16S rRNA gene ML tree (CCT) of <i>Alphaproteobacteria</i>	39
3	Unconstrained 16S rRNA gene ML tree (UCT) of <i>Alphaproteobacteria</i>	60
4	Supermatrix-based <i>Labrenzia</i> ML tree	61
5	Supermatrix-based <i>Sphingomonadales</i> ML tree	62

Abbreviations

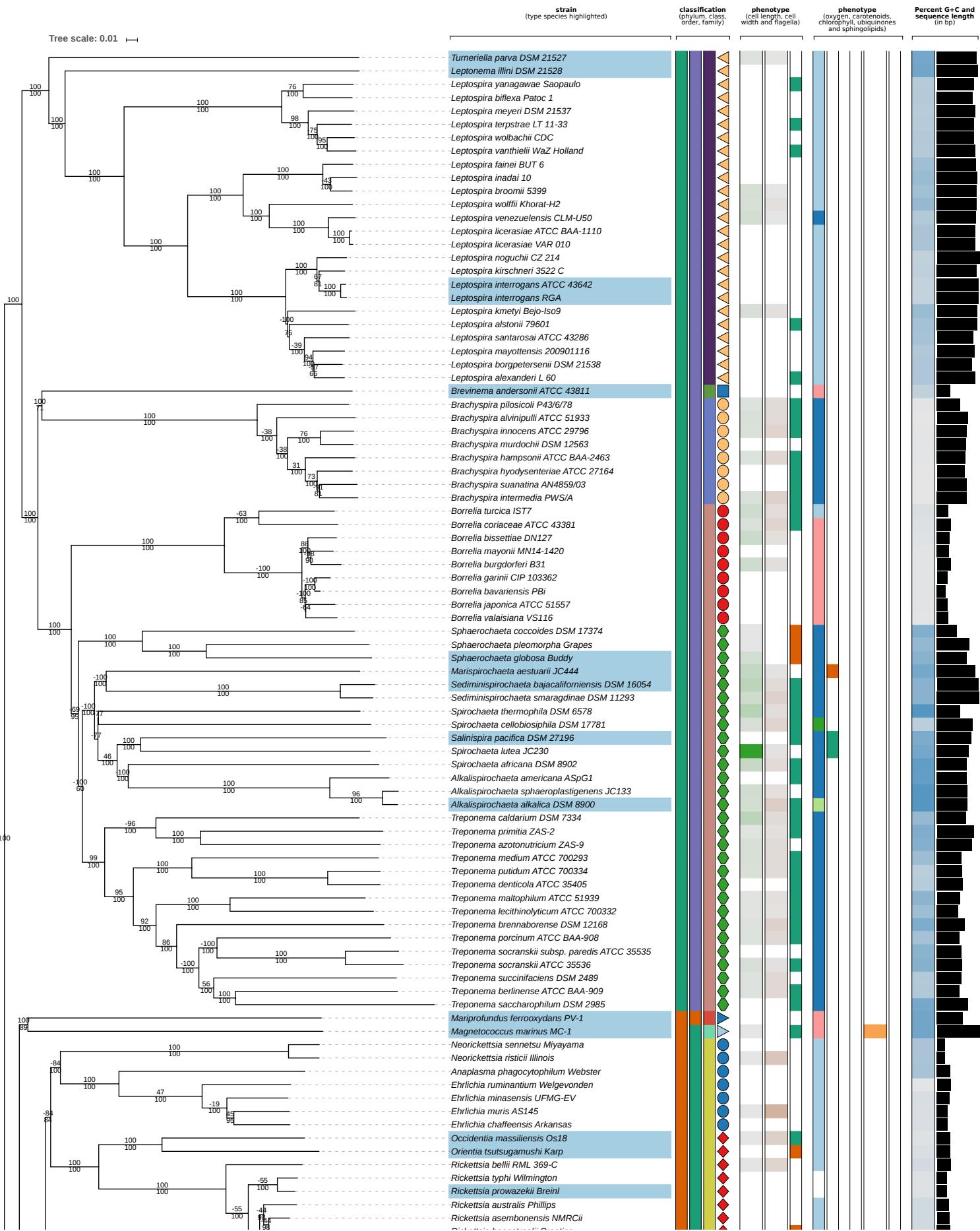
CCT constrained, comprehensive tree inferred with ML and MP using the bipartitions of the GBDP tree with $\geq 95\%$ support as backbone constraint

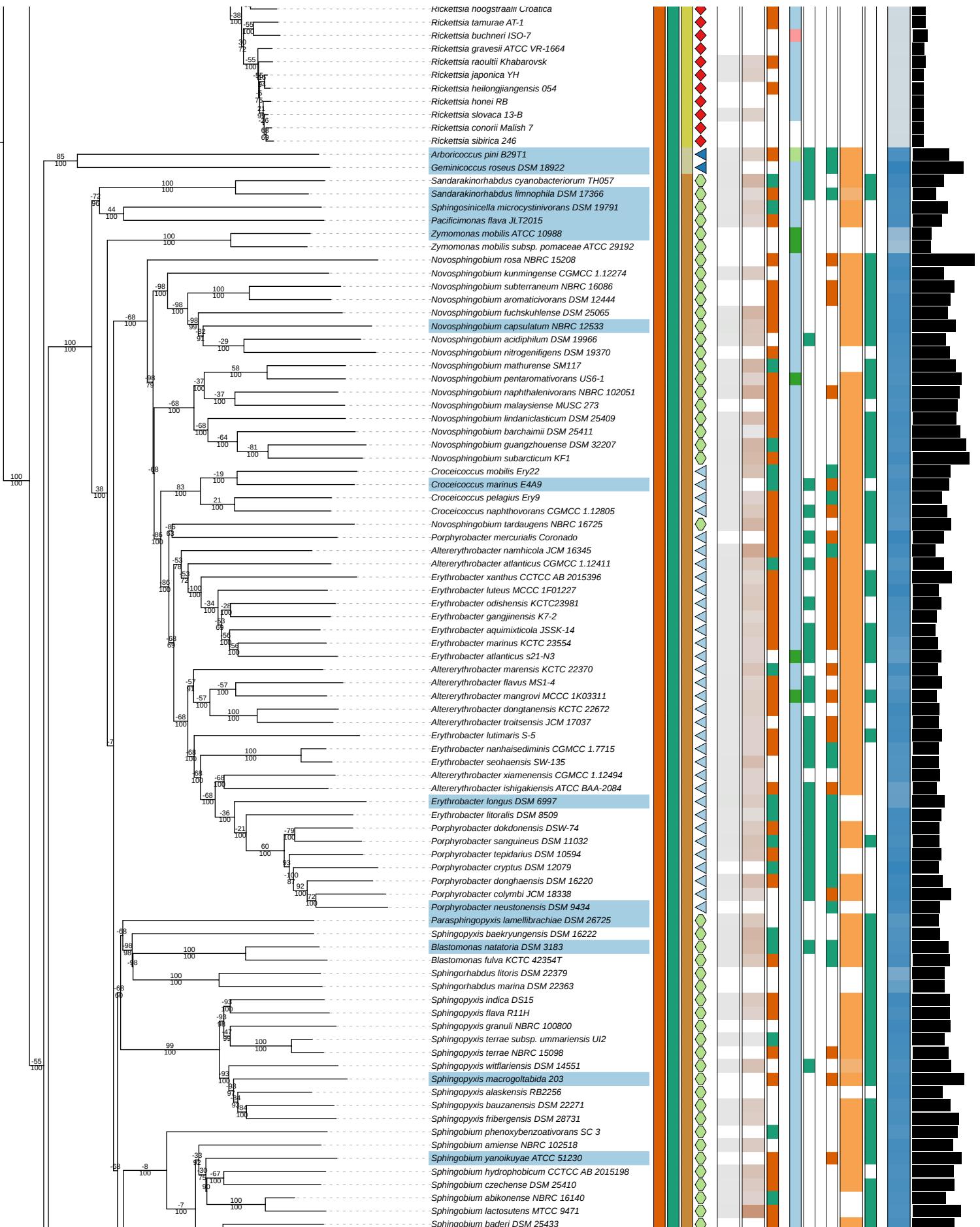
GBDP Genome BLAST Distance Phylogeny

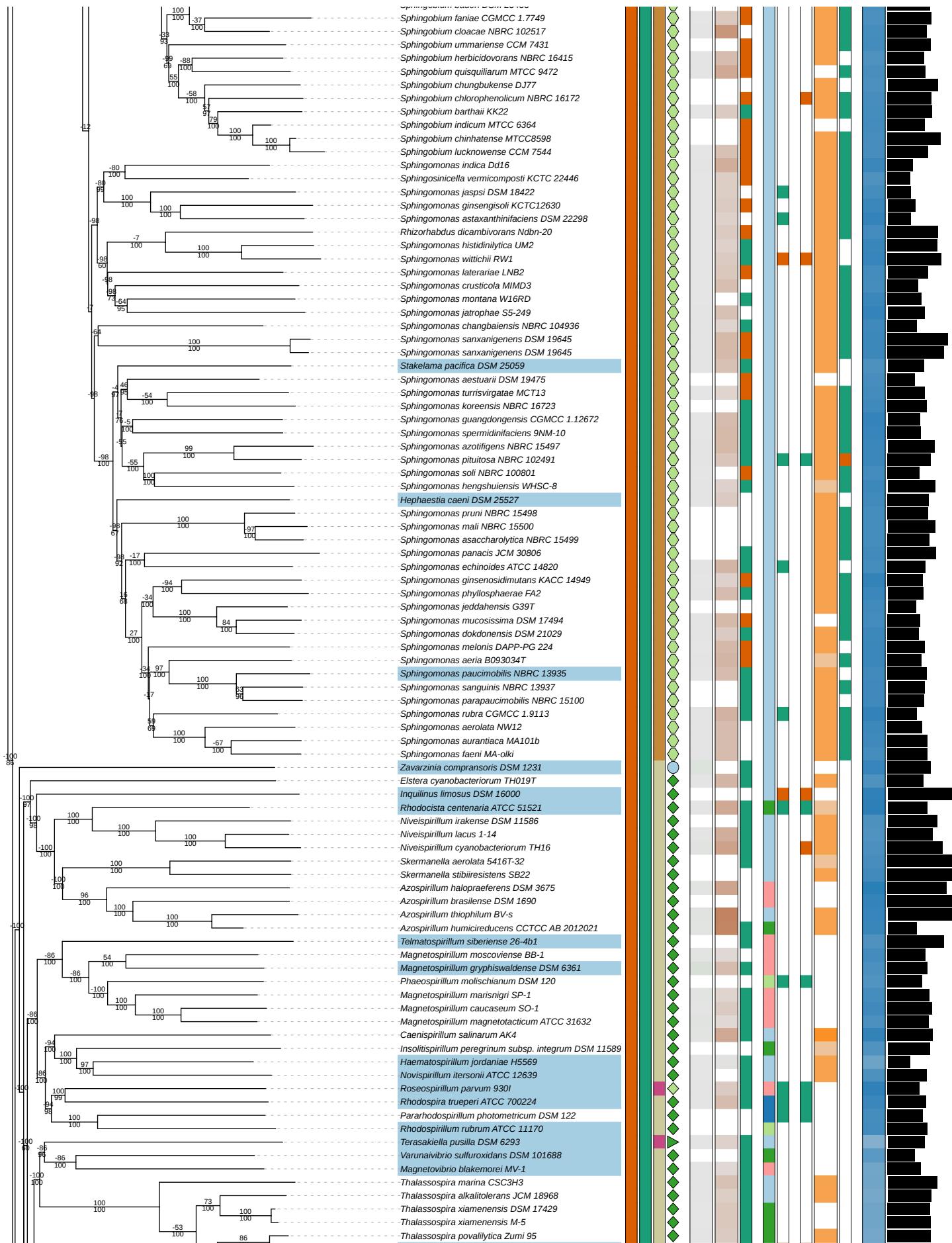
ML Maximum Likelihood

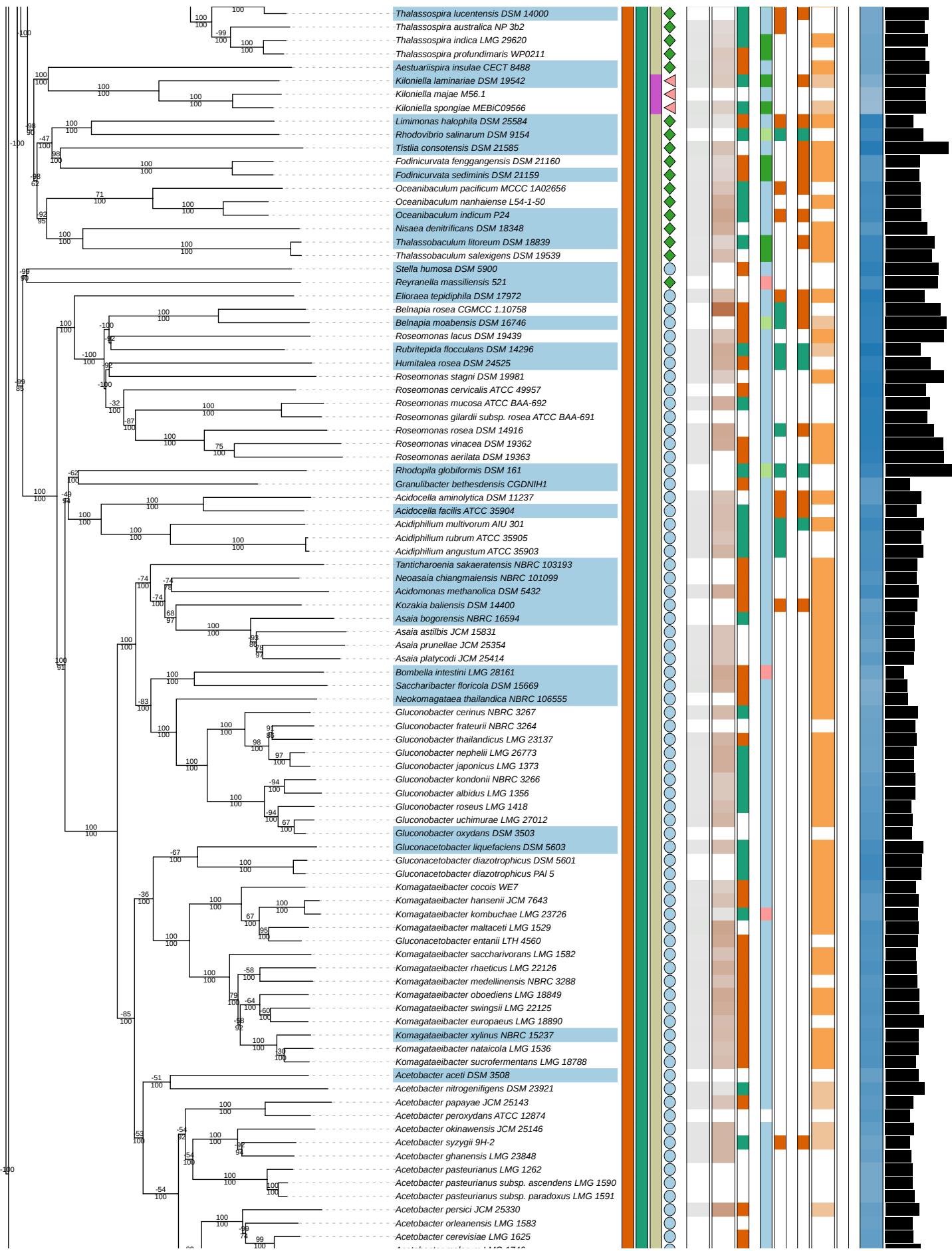
UCT unconstrained, comprehensive 16S rRNA gene tree

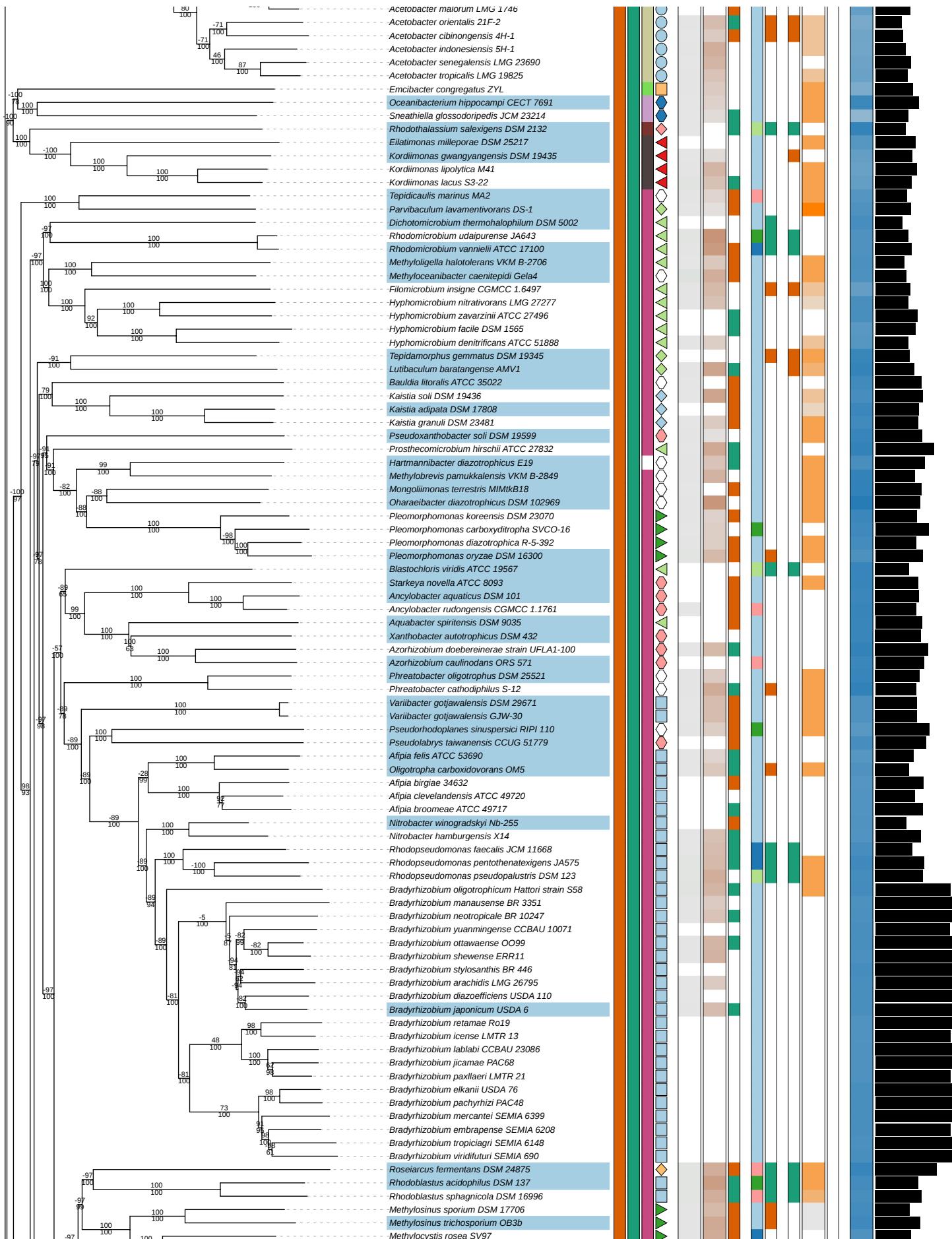


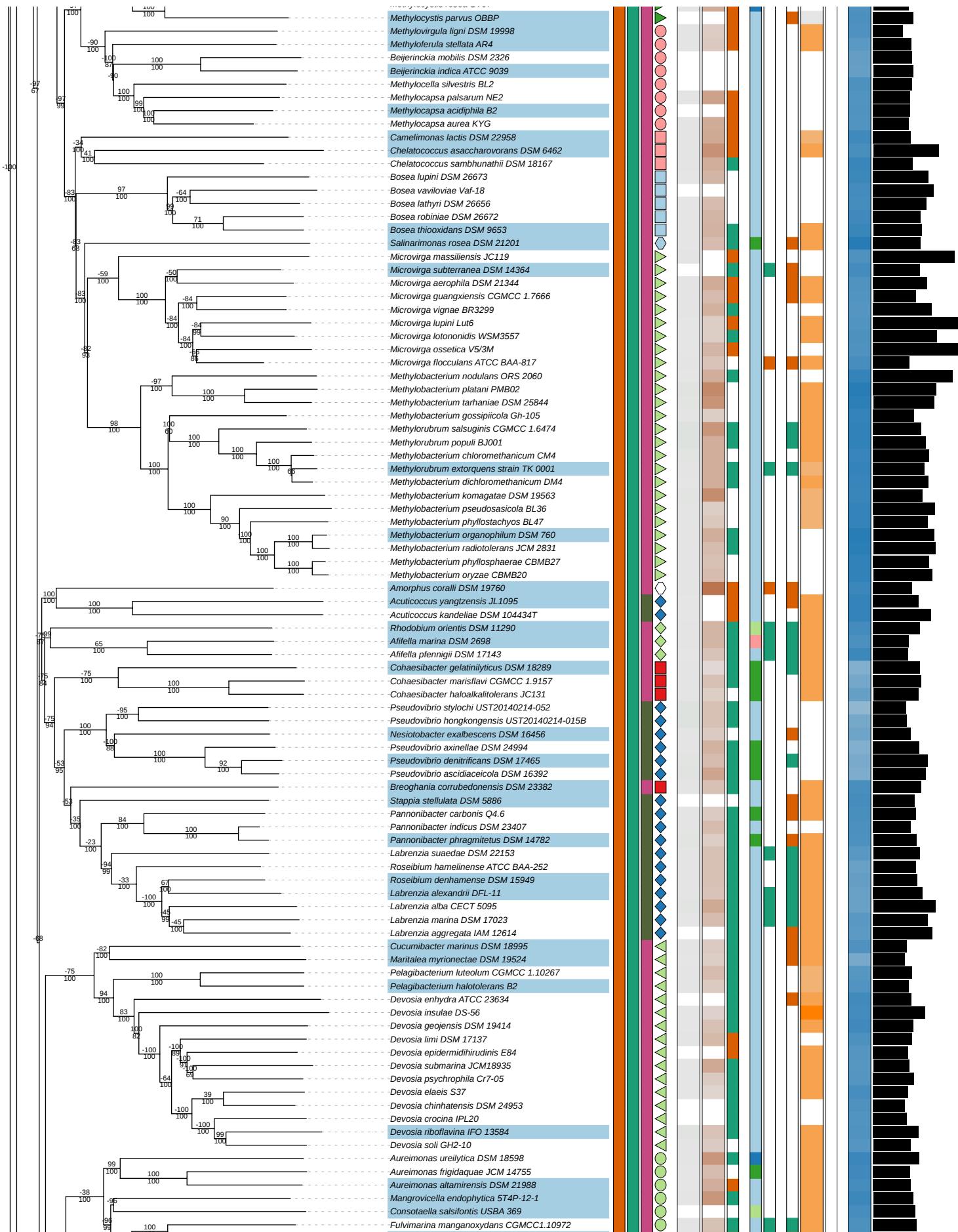


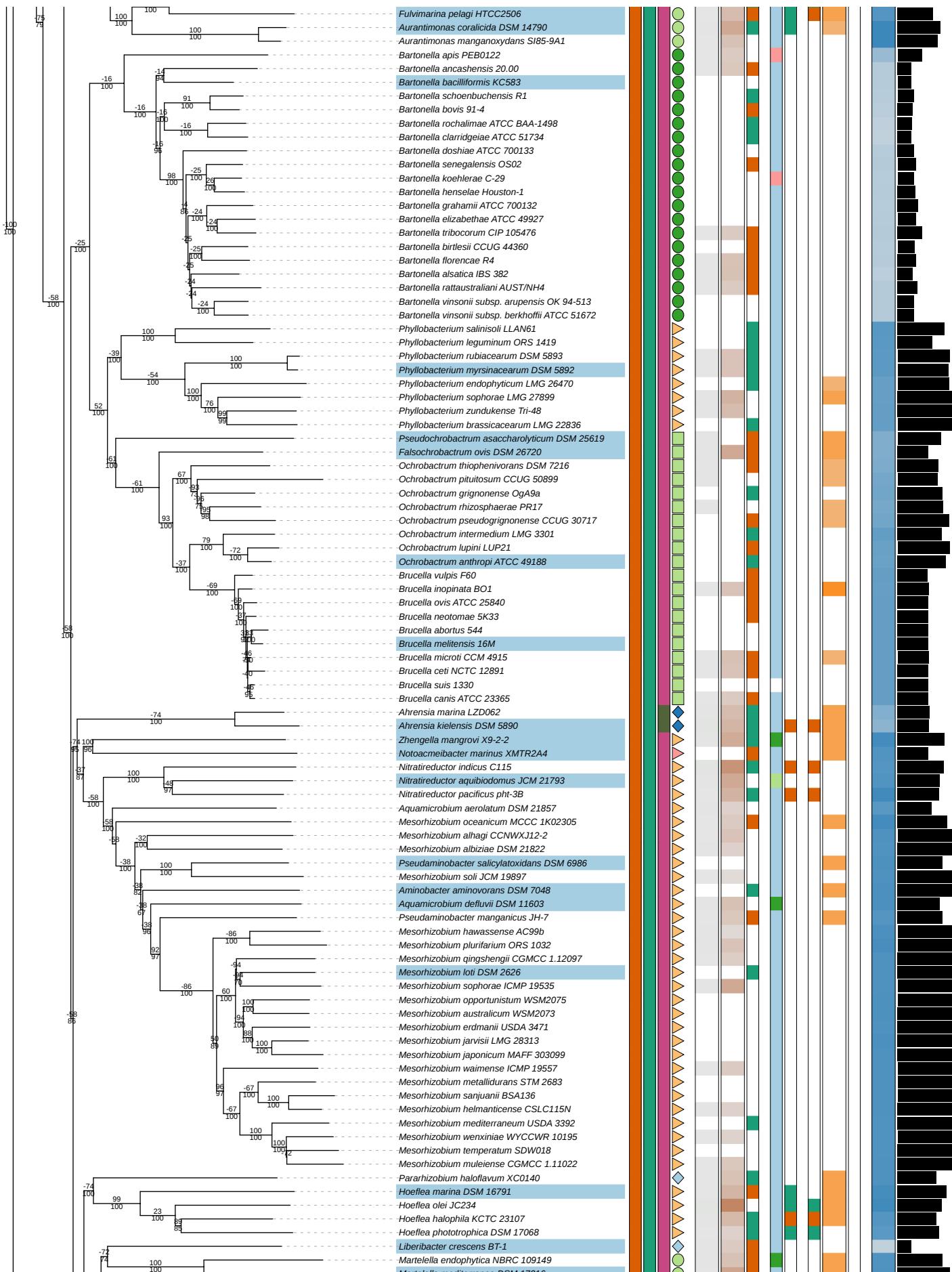


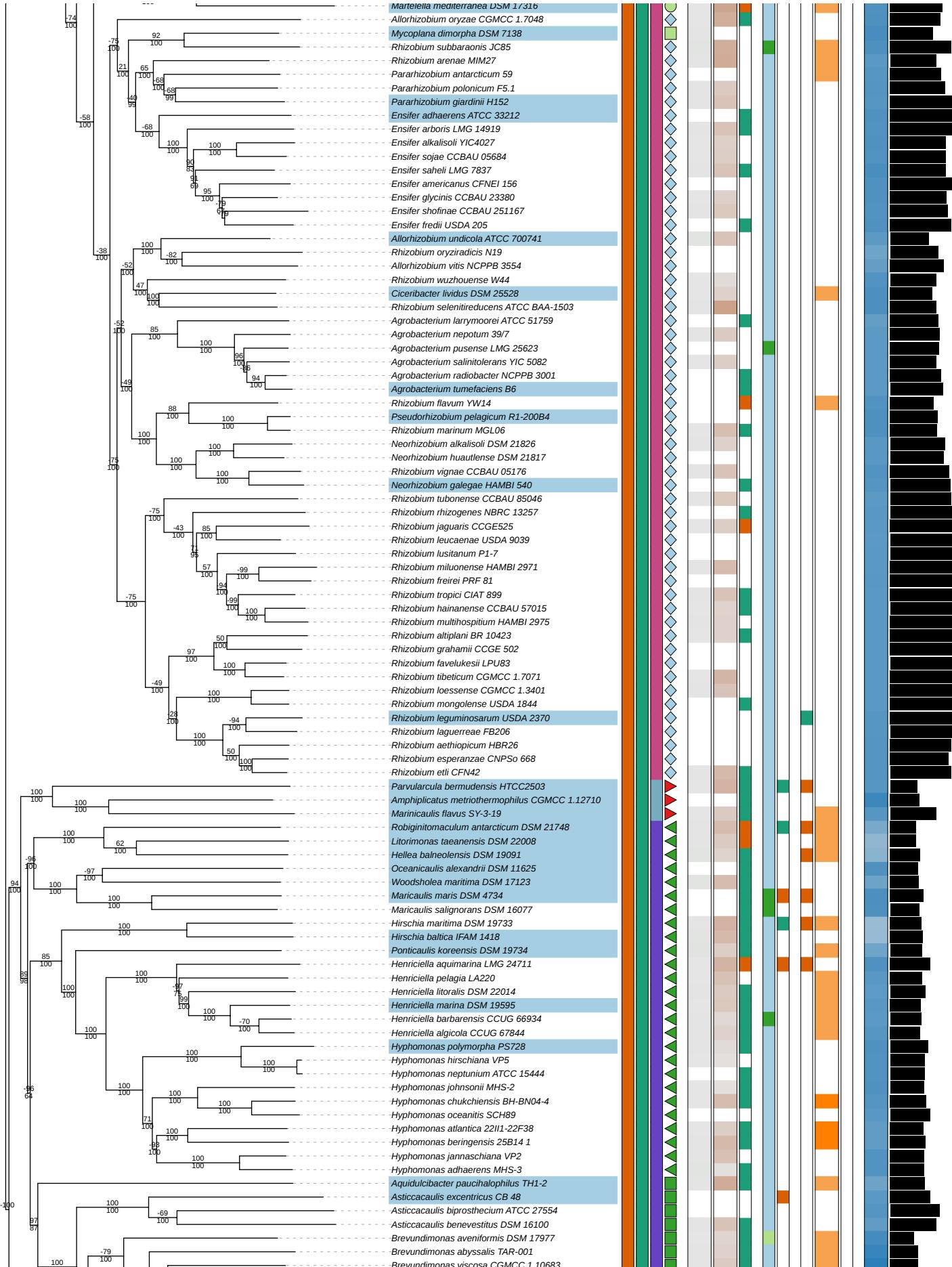


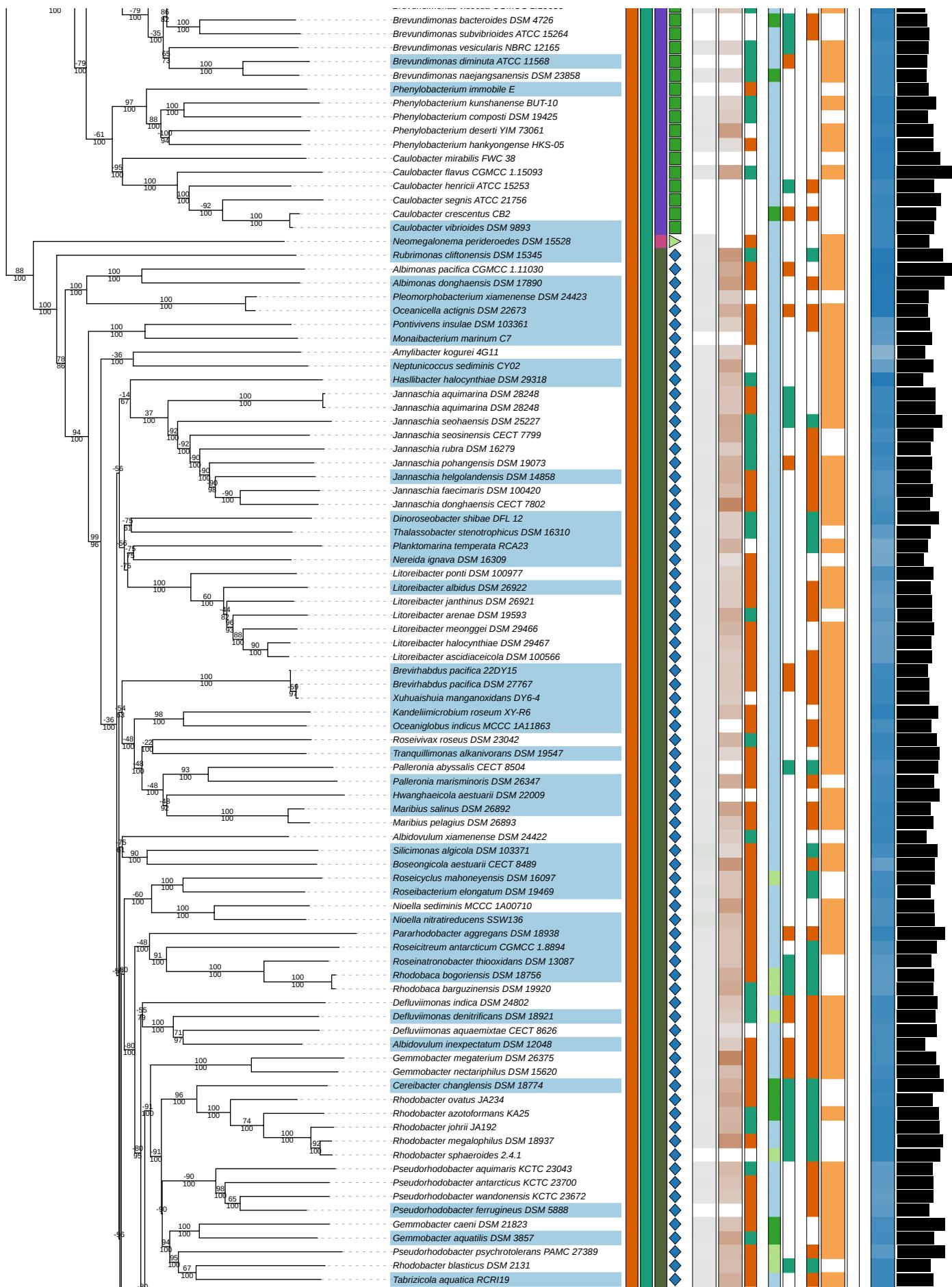


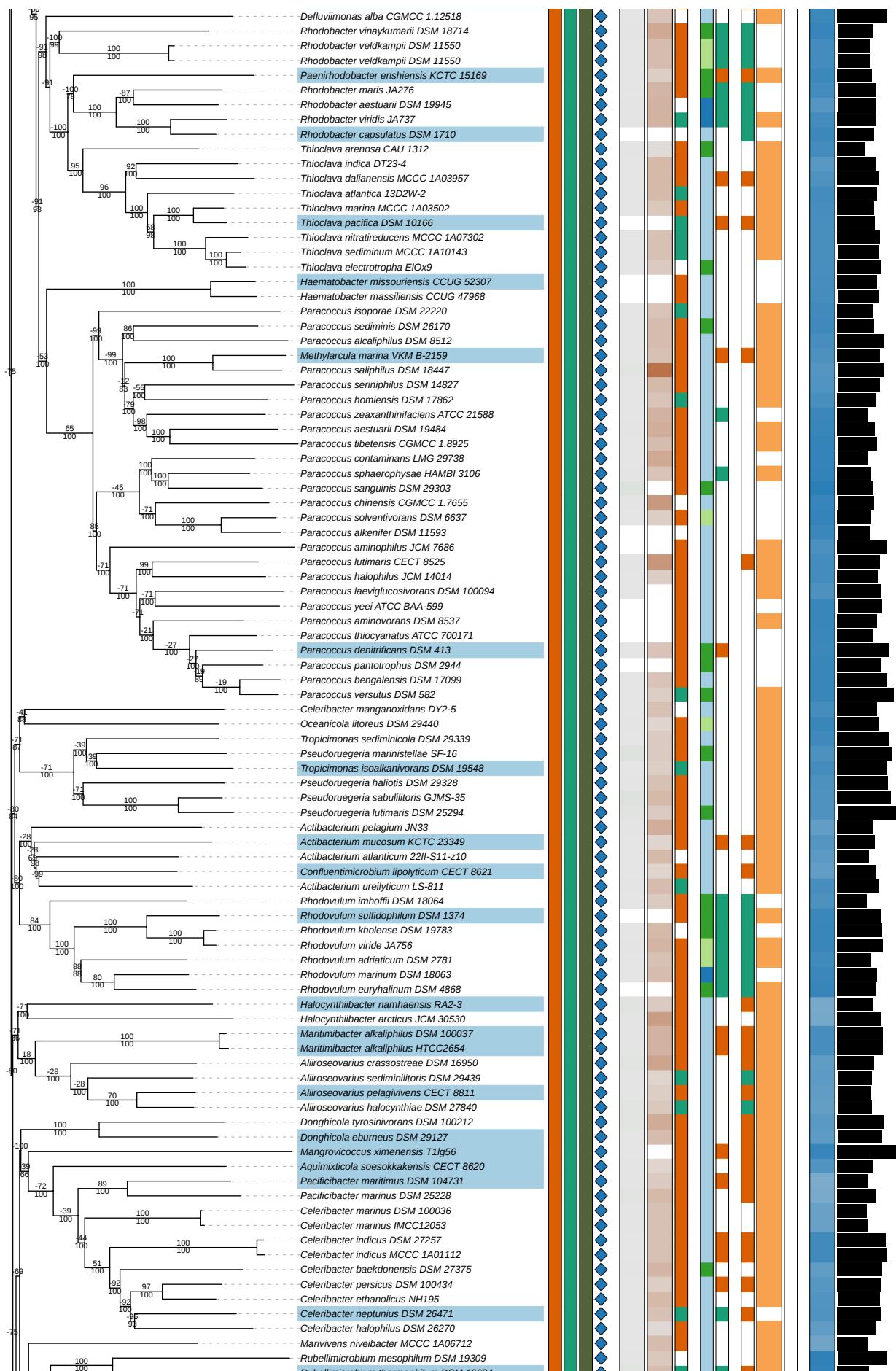


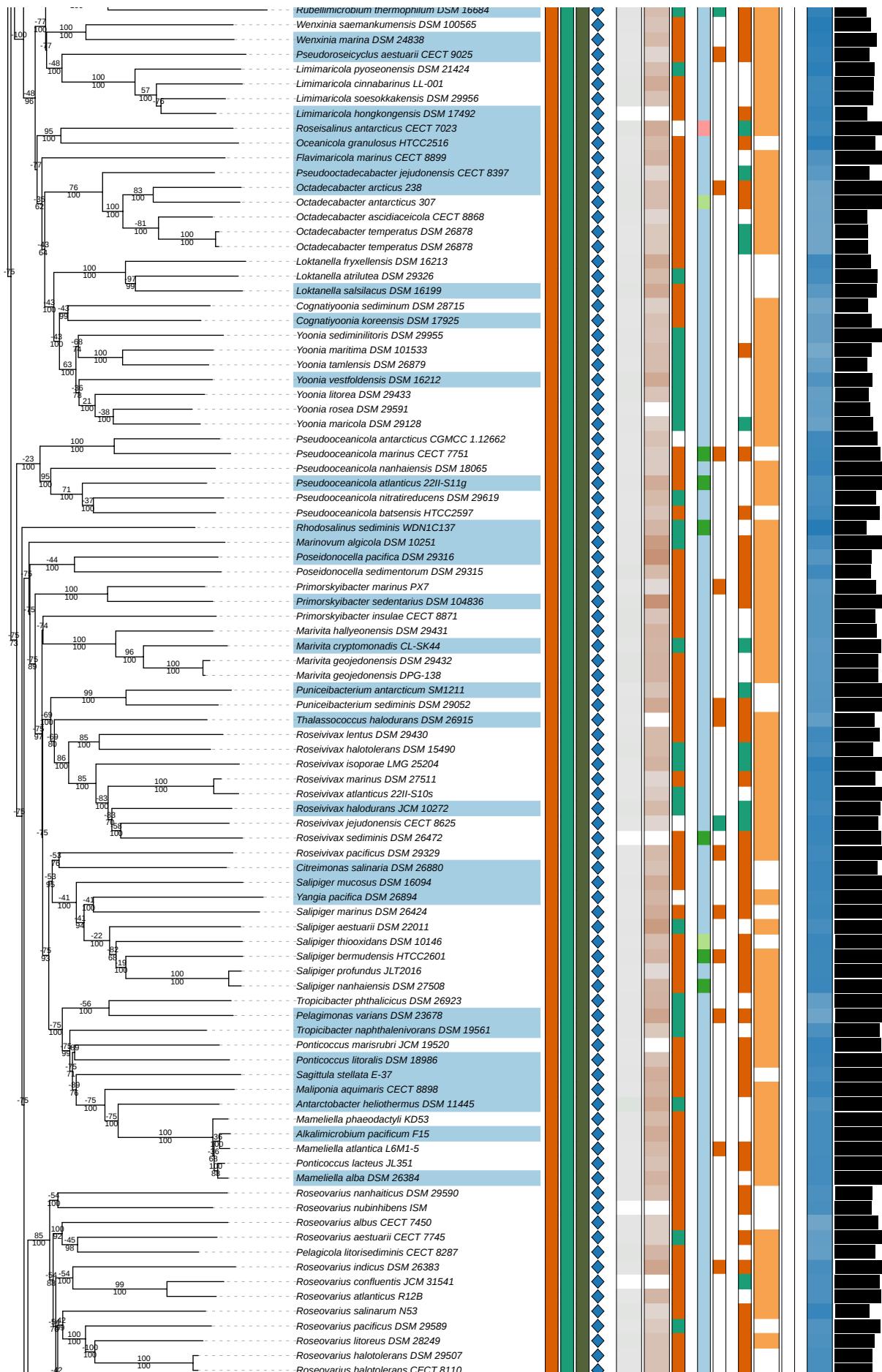












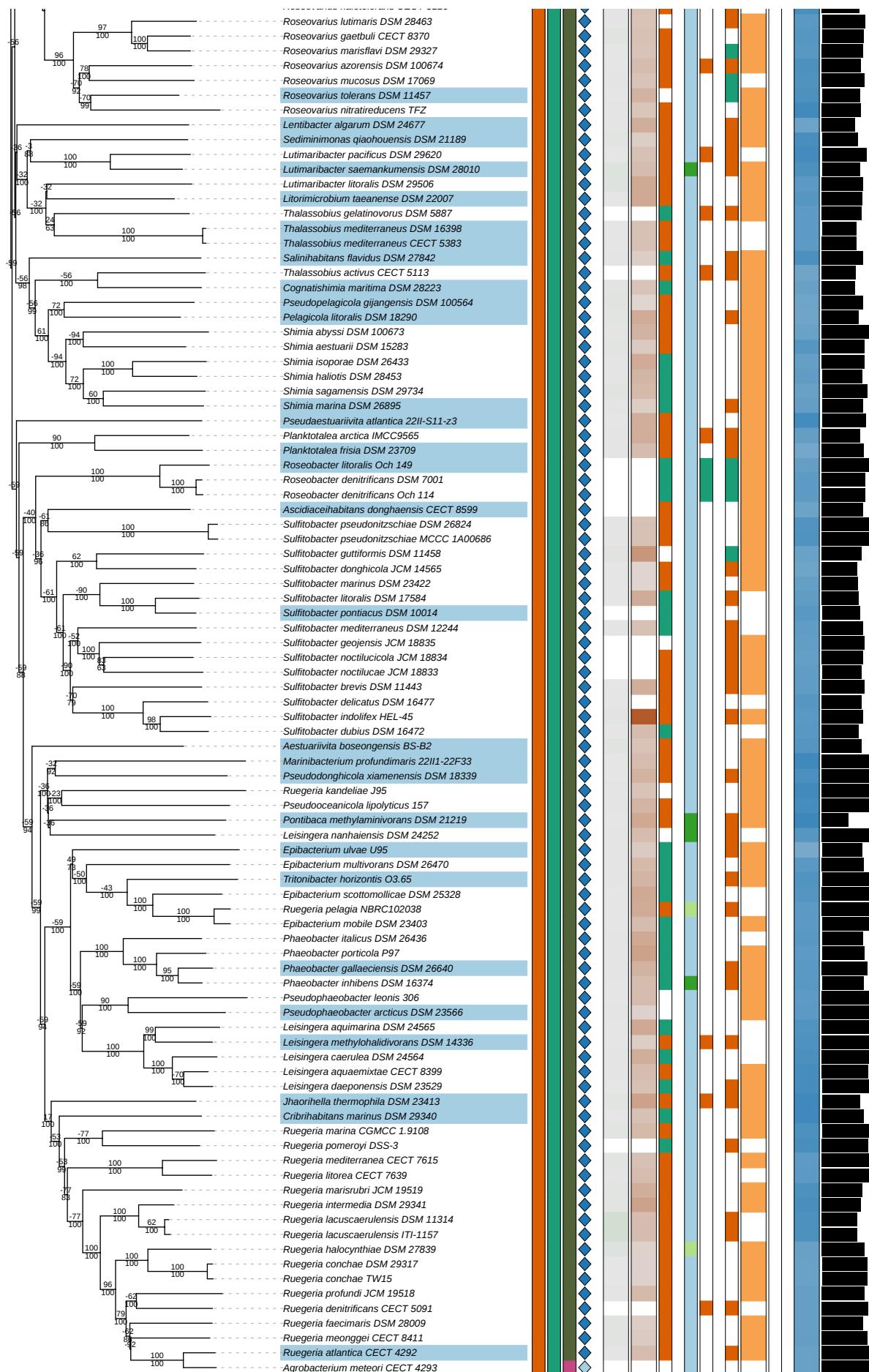
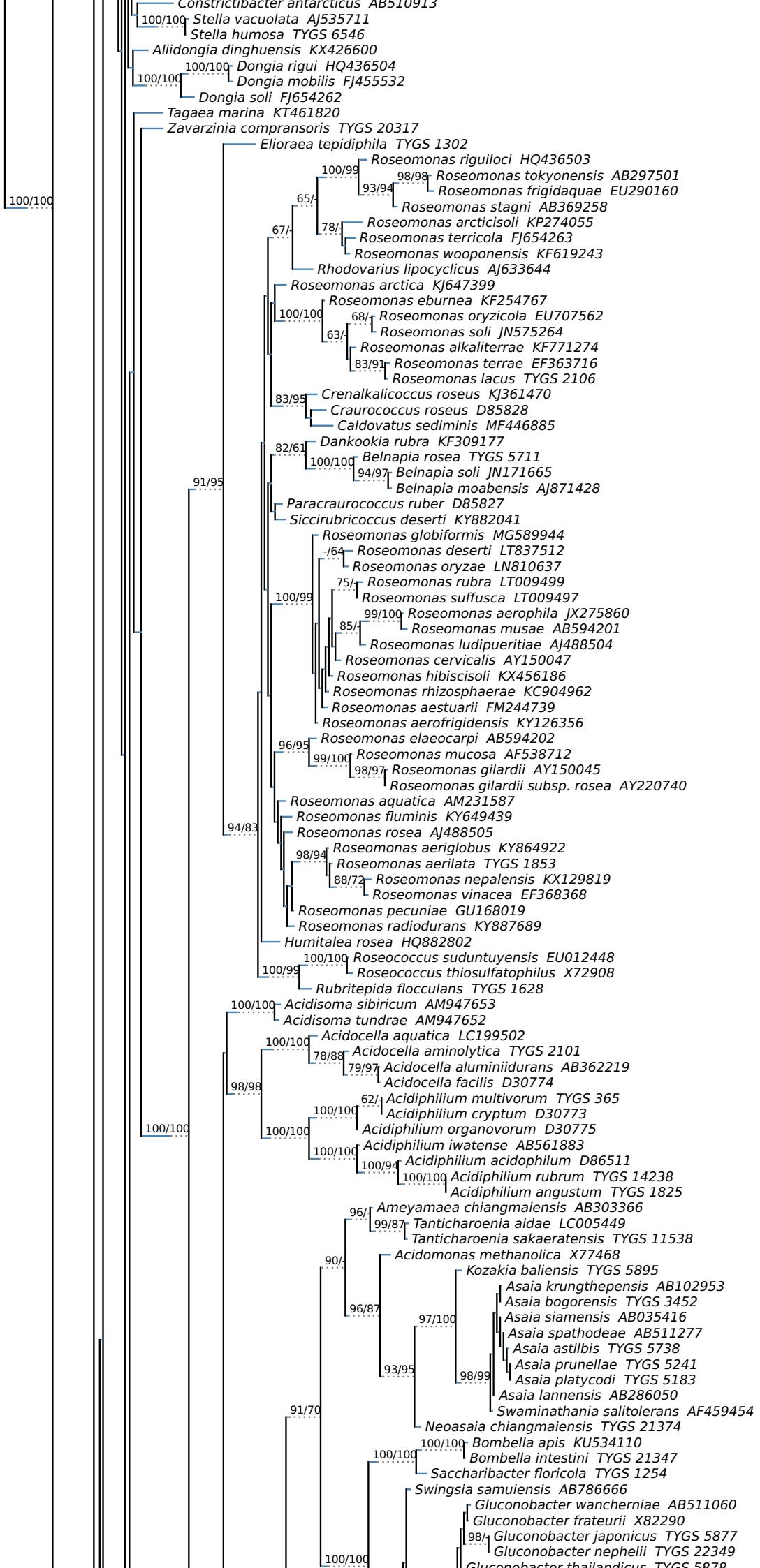
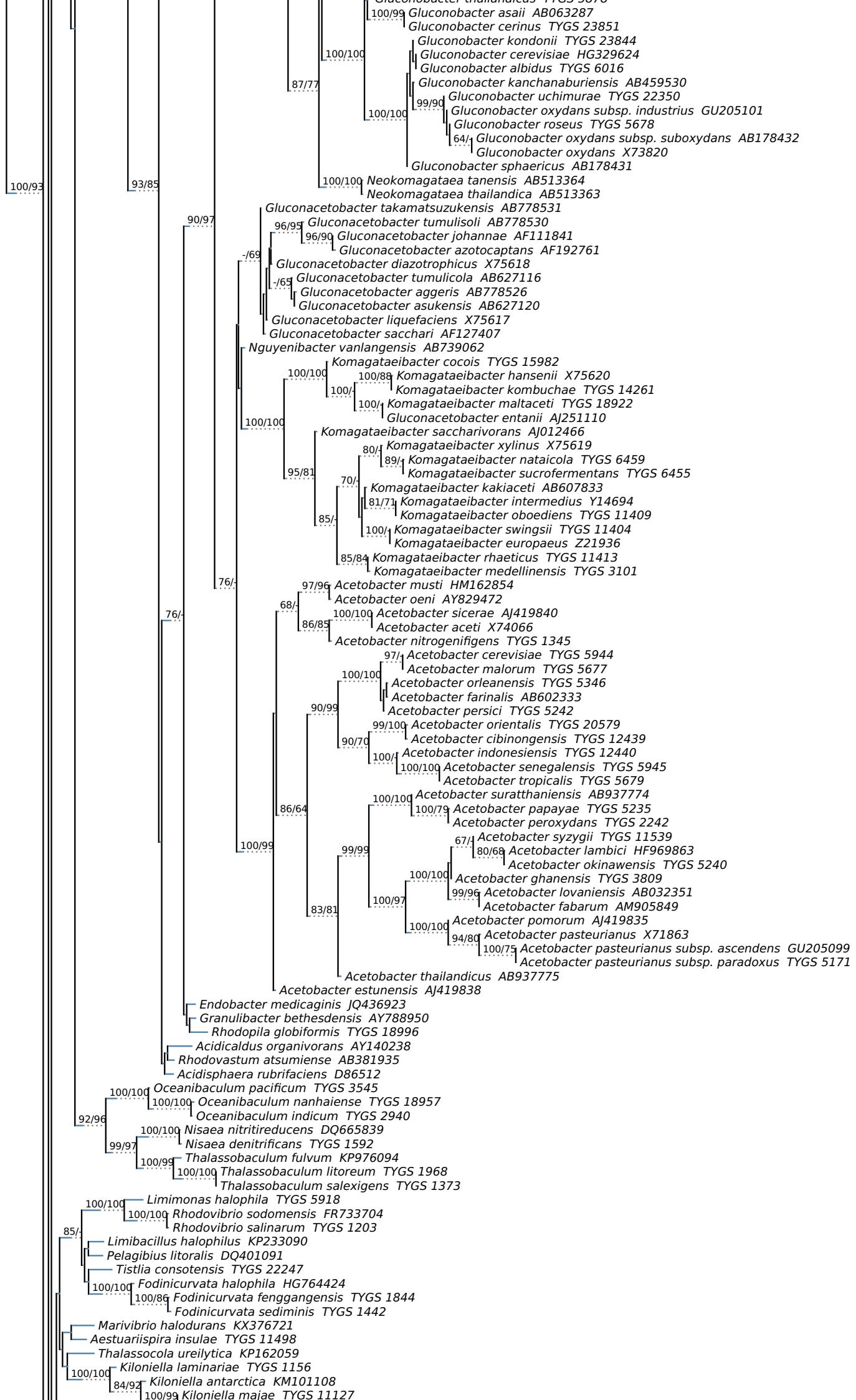
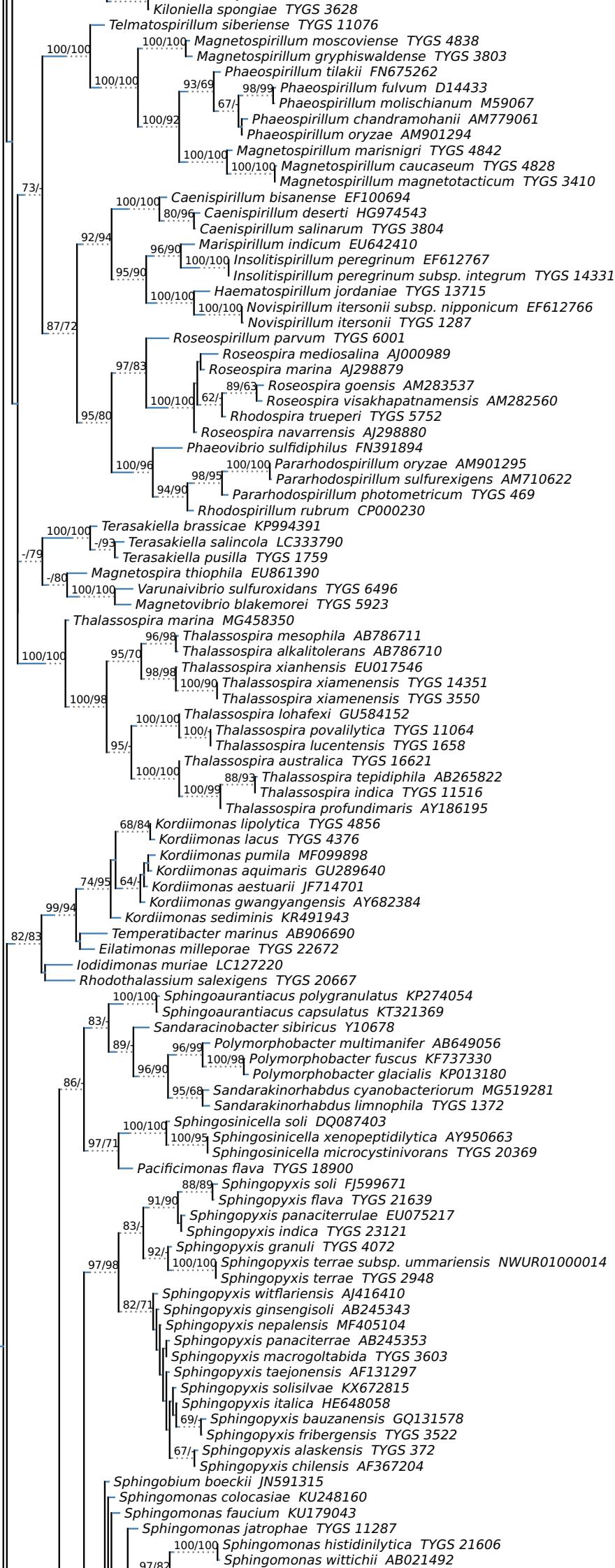


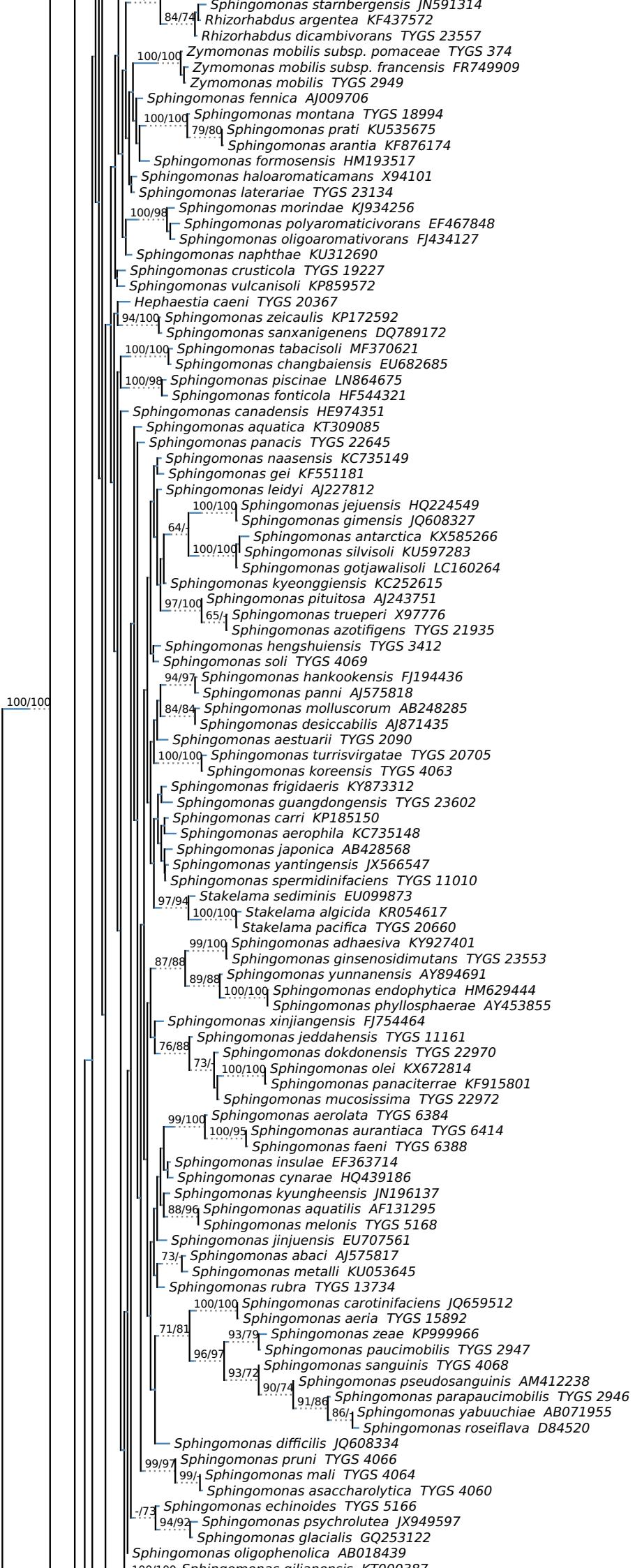
Figure 1: The figure shows the complete, uncollapsed GBDP tree of the *Alphaproteobacteria* genome dataset, which had to be distributed across Figures 1-9 in the main manuscript. Tree inferred with FastME from GBDP distances calculated from whole proteomes. The branches are scaled in terms of log-transformed intergenomic distances (GBDP formula d_5). The numbers below branches are GBDP pseudo-bootstrap support values from 100 replications. The numbers above branches are support (positive) or conflict (negative) values from the gene-content analysis. Tip colors indicate type species, colors to the right of the tips indicate, from left to right, phylum (1), class (2), order (3) and family (4). The blocks labelled as "phenotype" display phenotypic information (5-13), whereas the blue gradient scale (14) indicates the exact G+C content as calculated from the genome sequences. Genome size (15) is displayed at the right-hand side. See the embedded legend for details. *Jannaschia aquimarina* DSM 28248 is represented by two distinct GenBank biosample accessions (above: SAMN05421775, below: SAMN03329626). The same holds true for *Octadecabacter temperatus* DSM 26878 (above: SAMN03891579, below: SAMN05444287), *Rhodobacter veldkampii* DSM 11550 (above: SAMN10866319, below: SAMN08535030) and *Sphingomonas sanxanigenens* DSM 19645 (above: SAMN02641489, below: SAMN02745820).

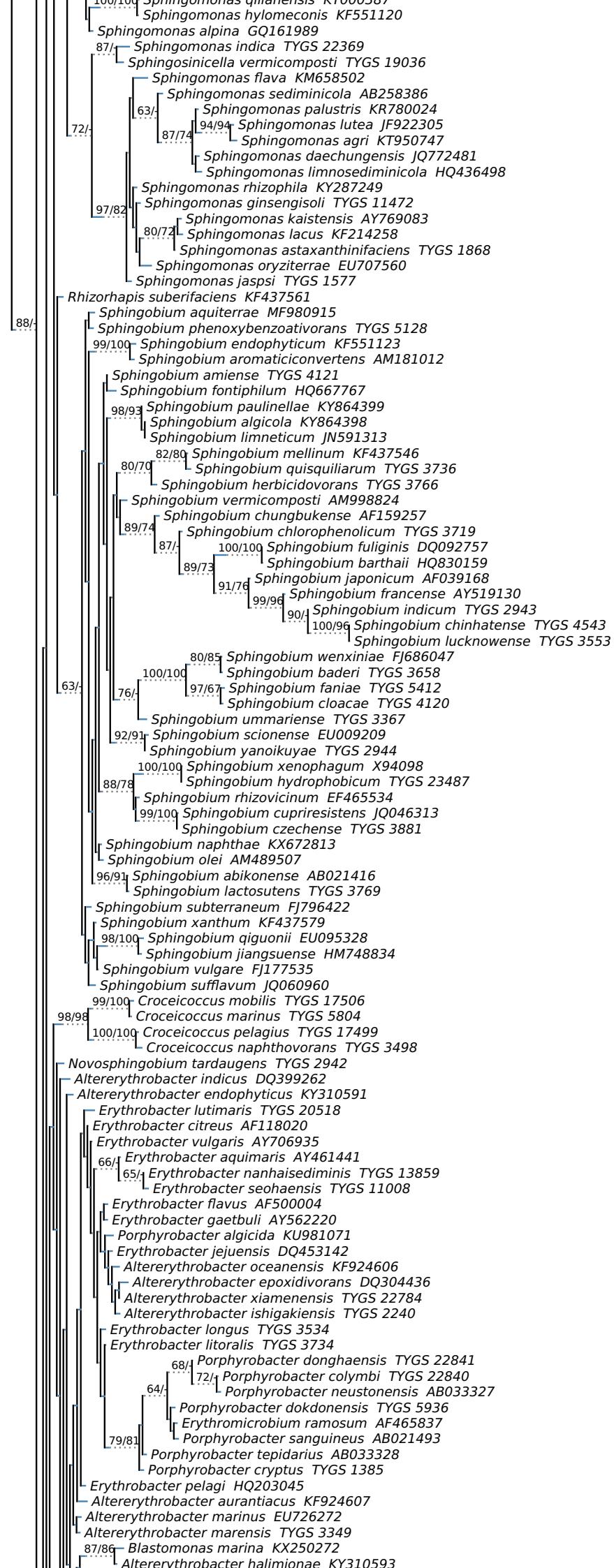


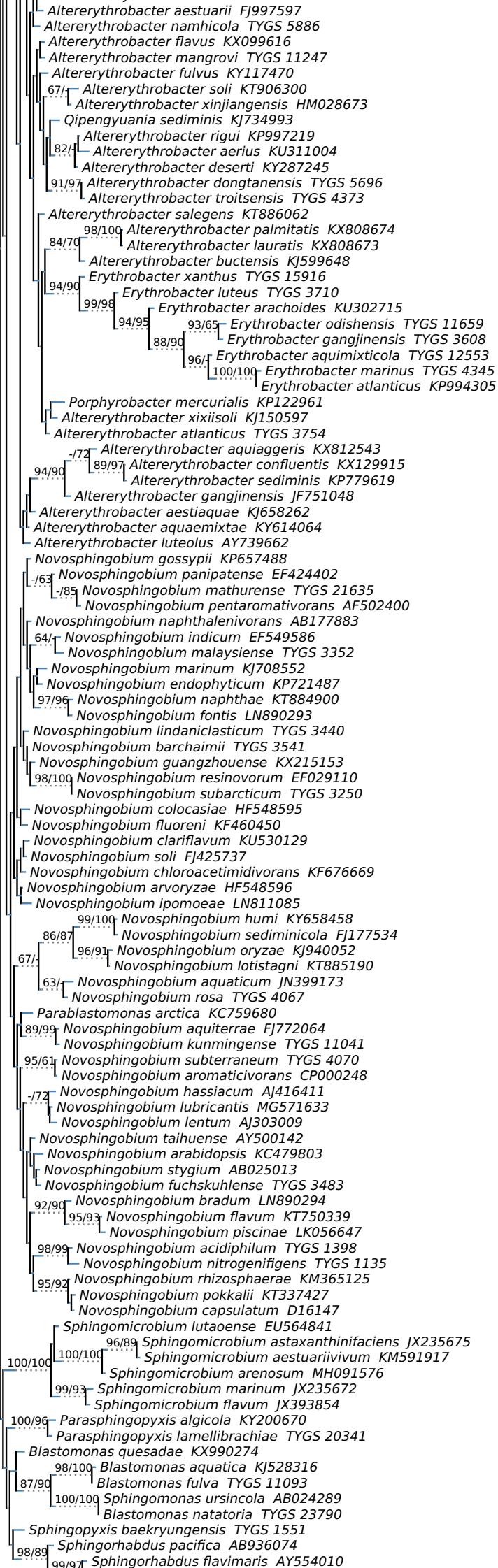


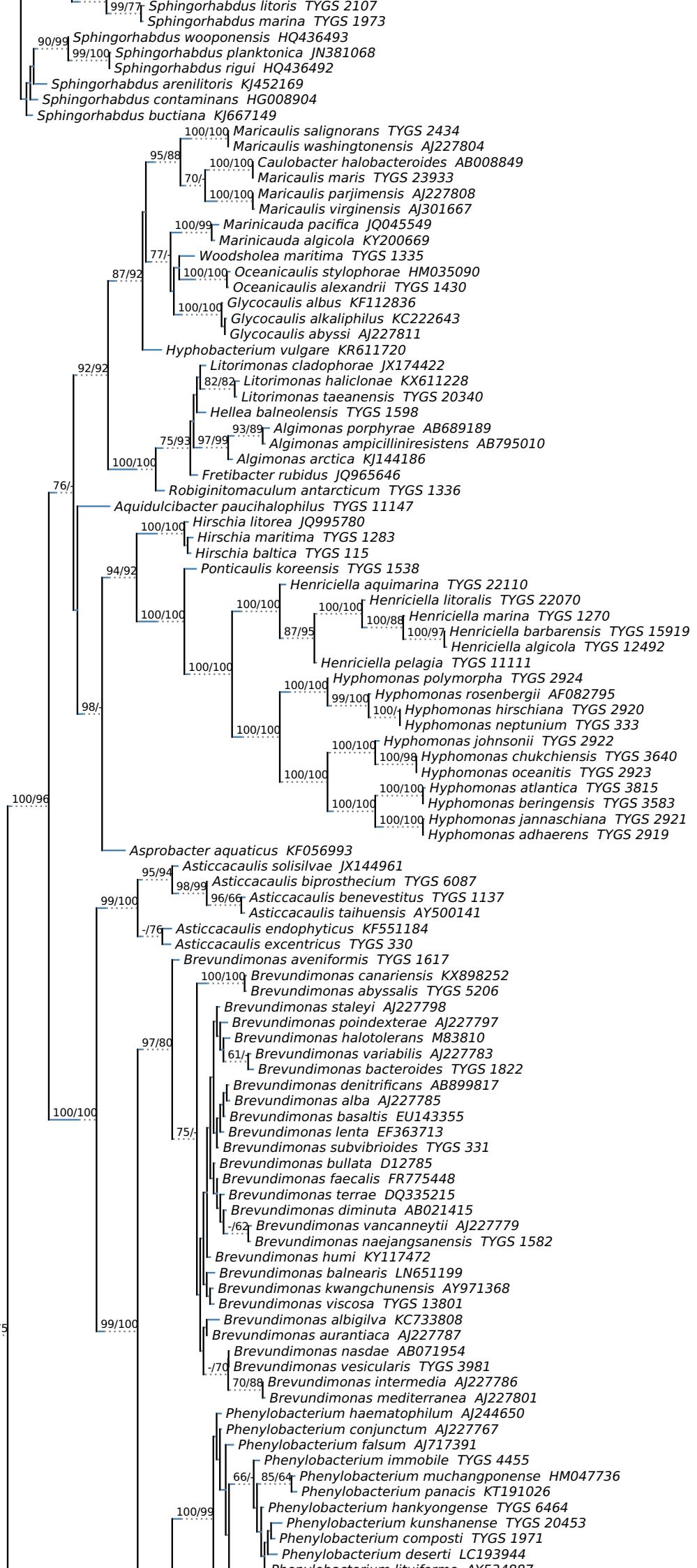


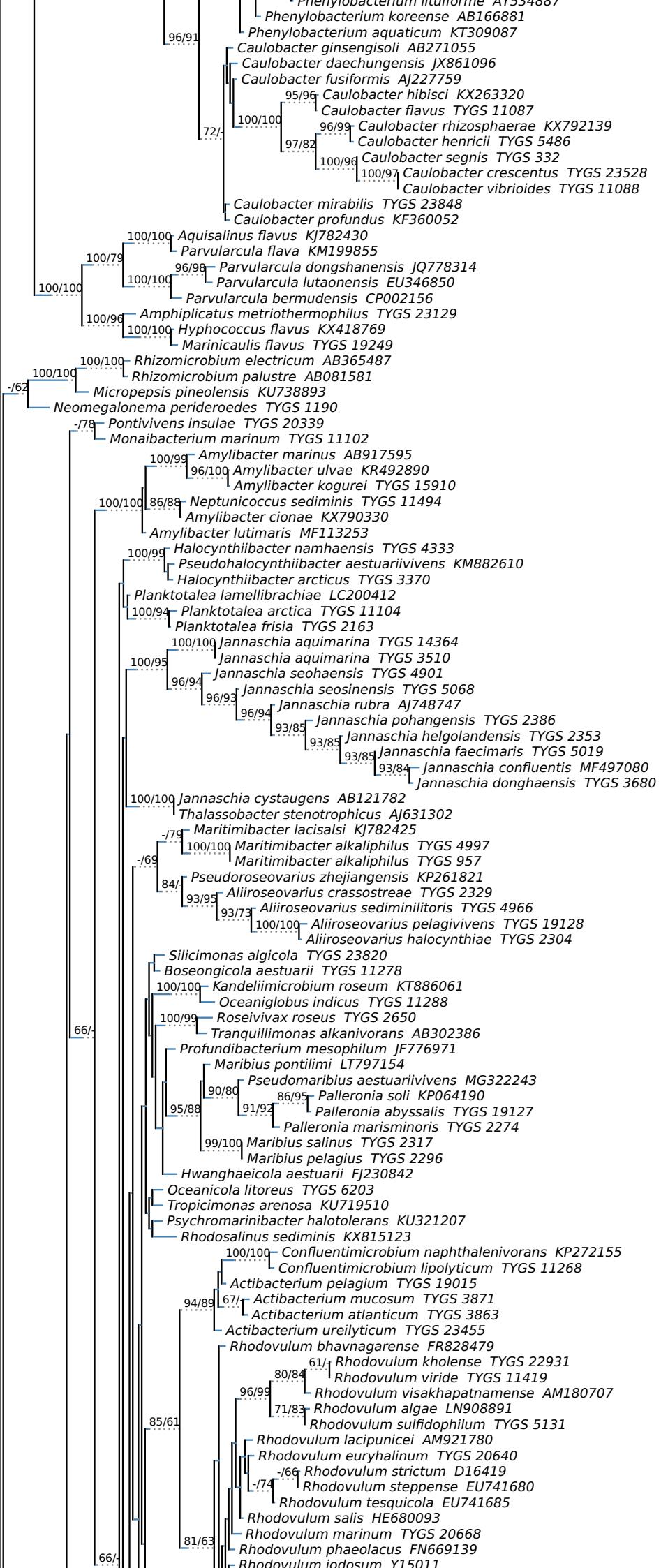


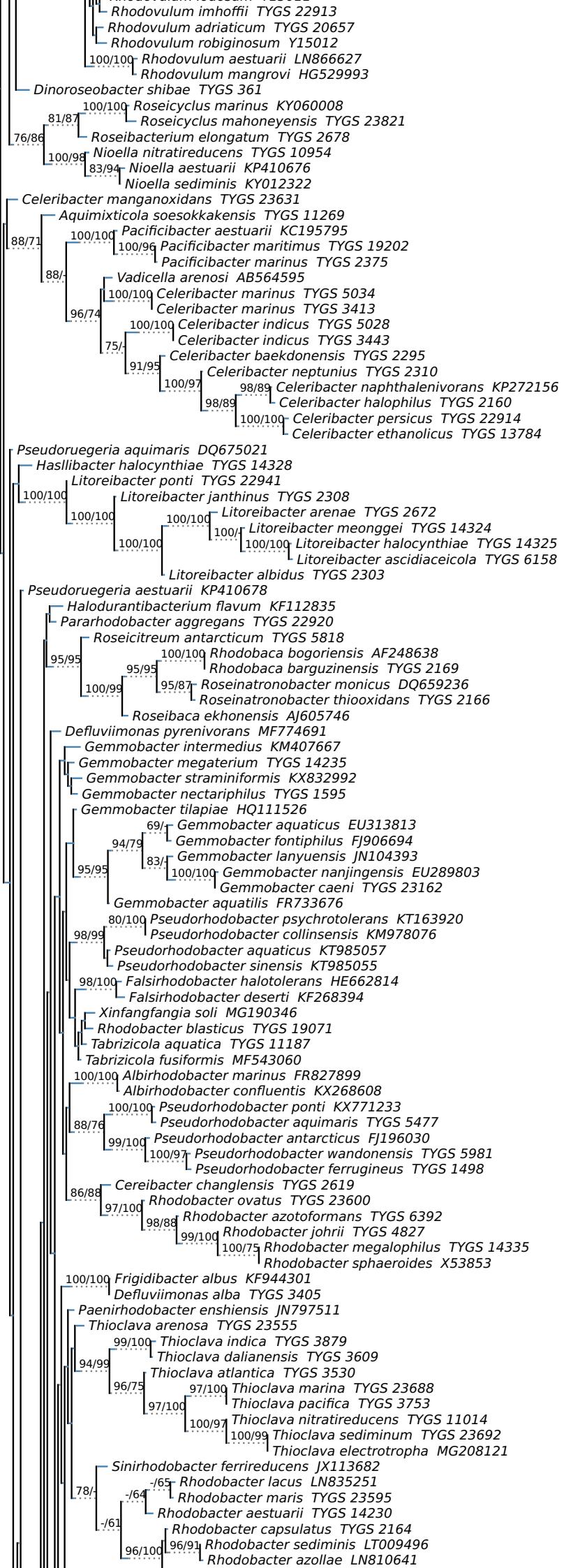




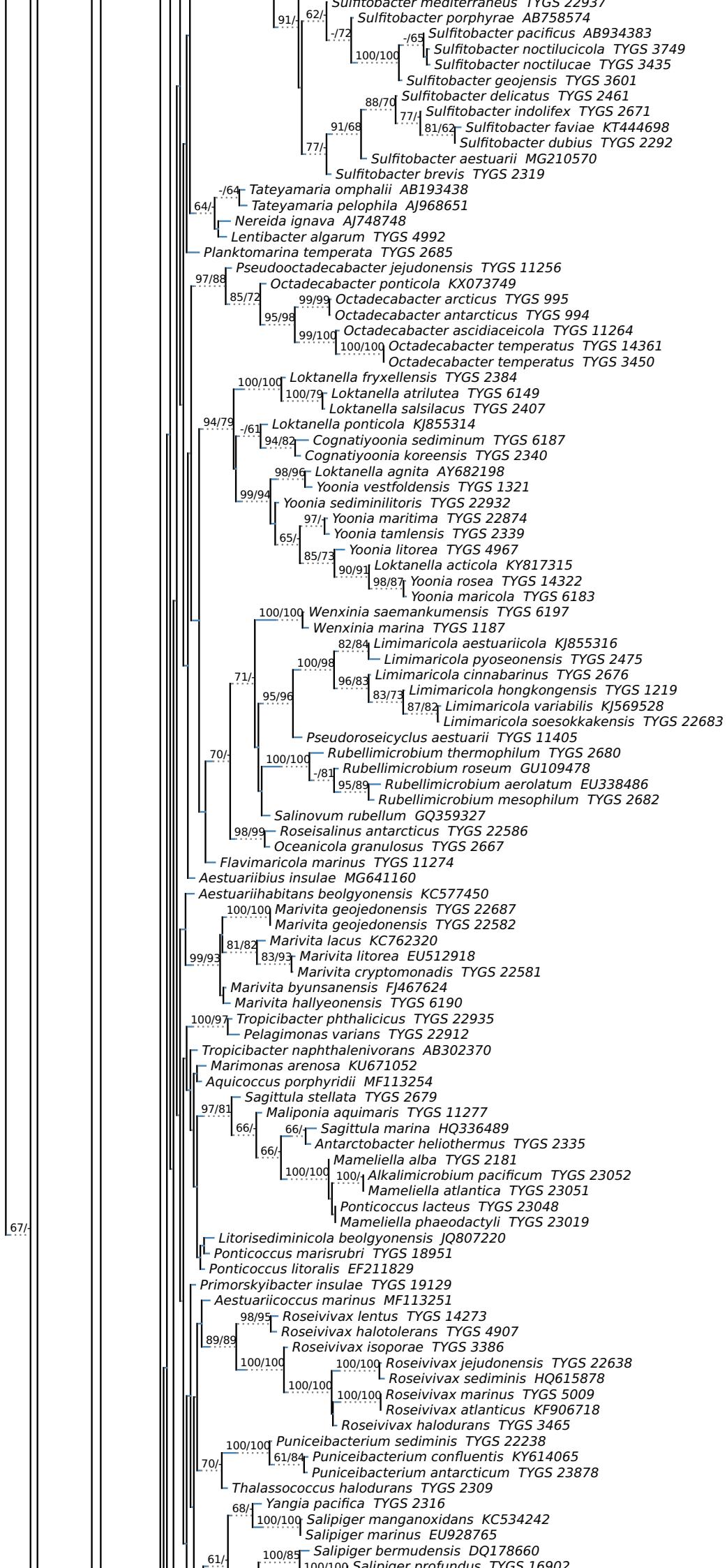


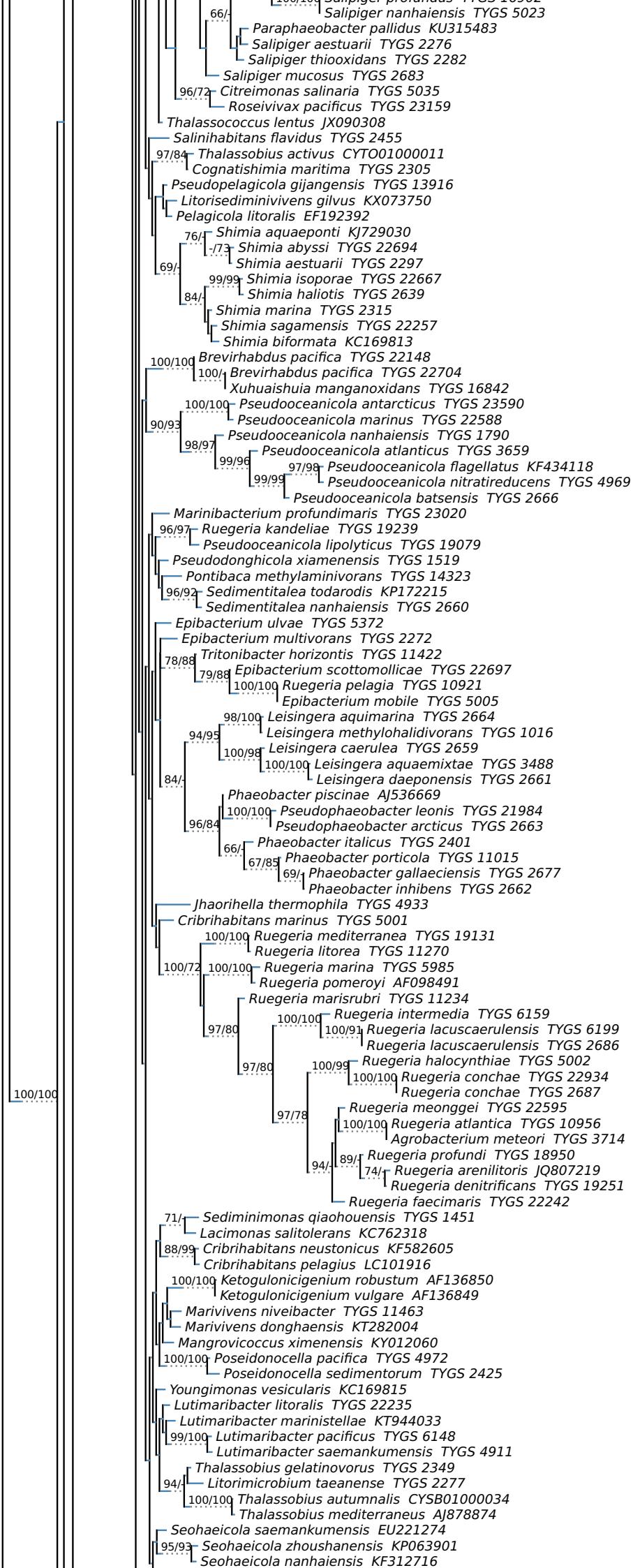


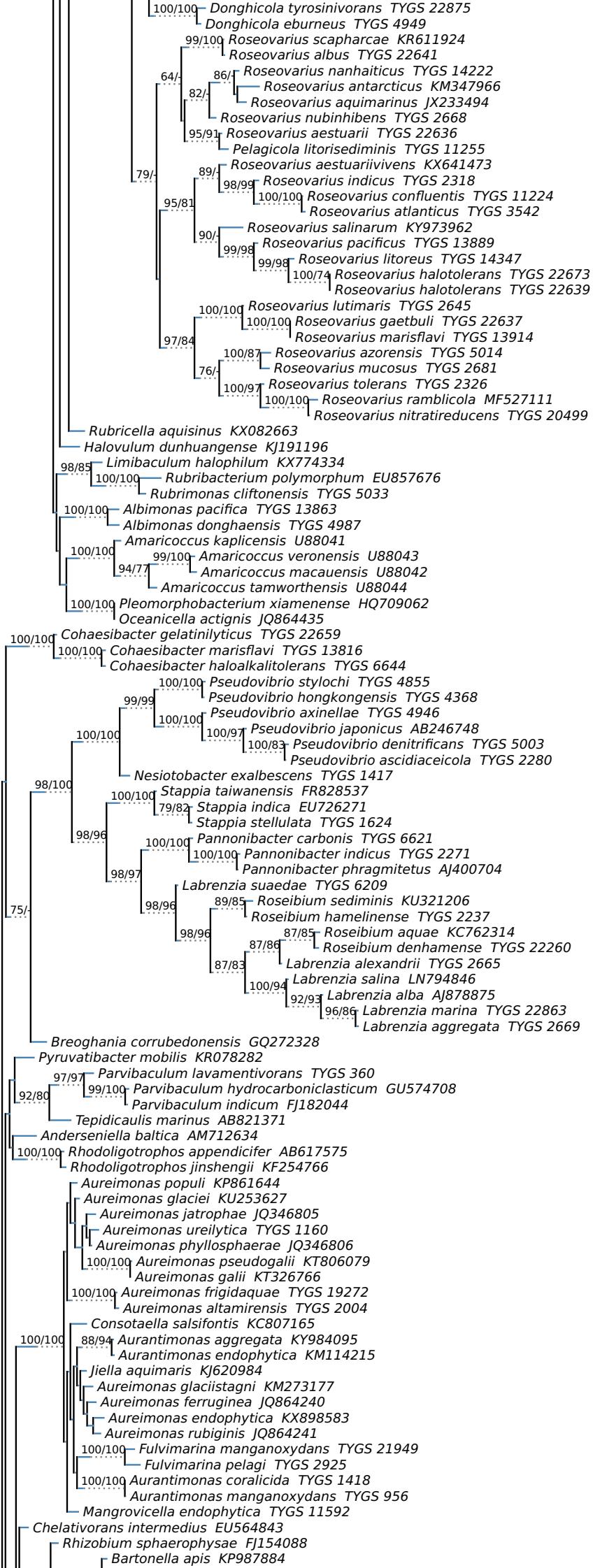


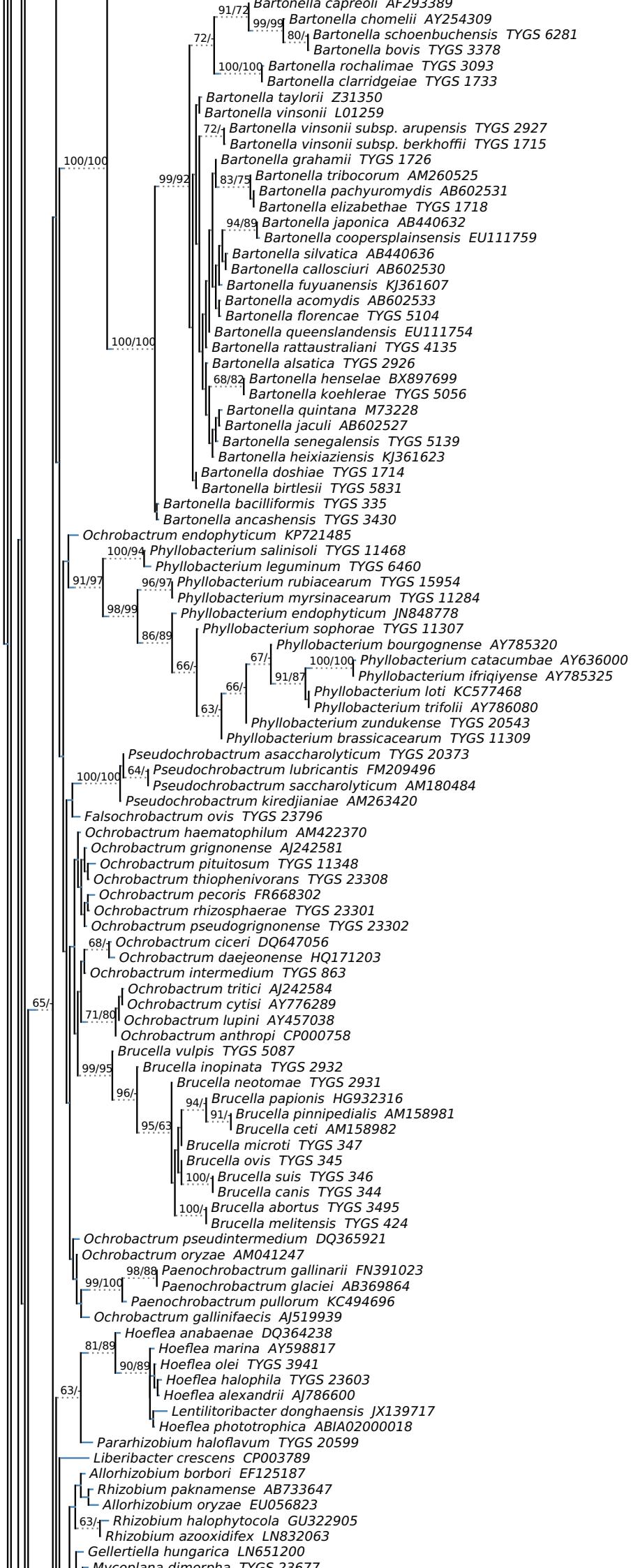


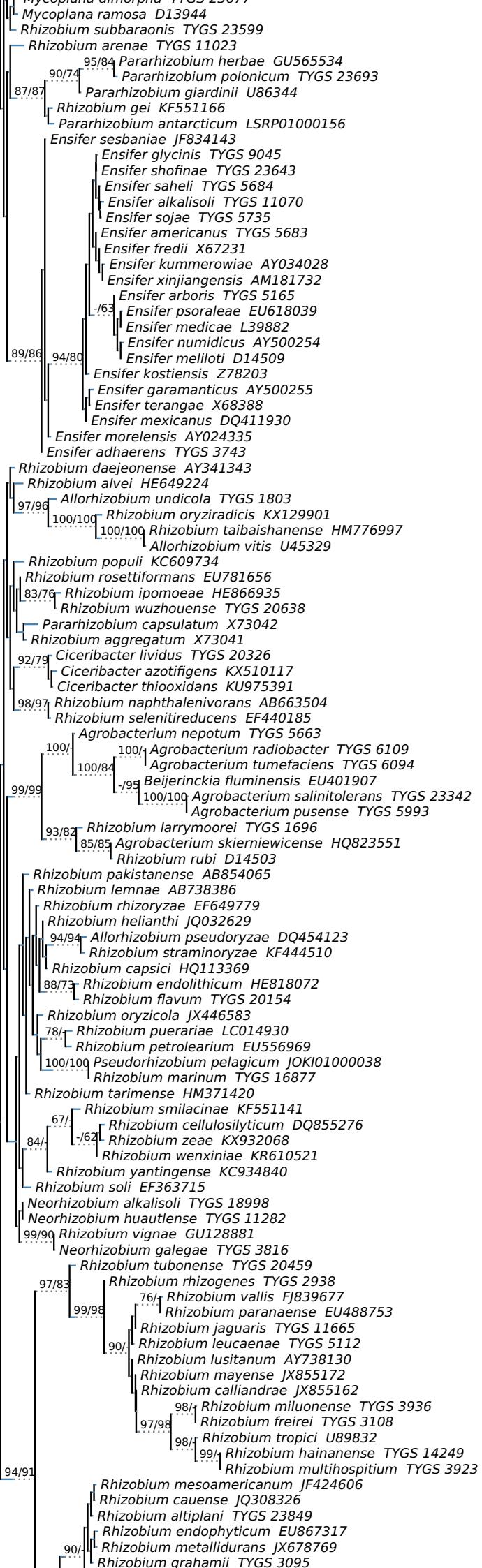


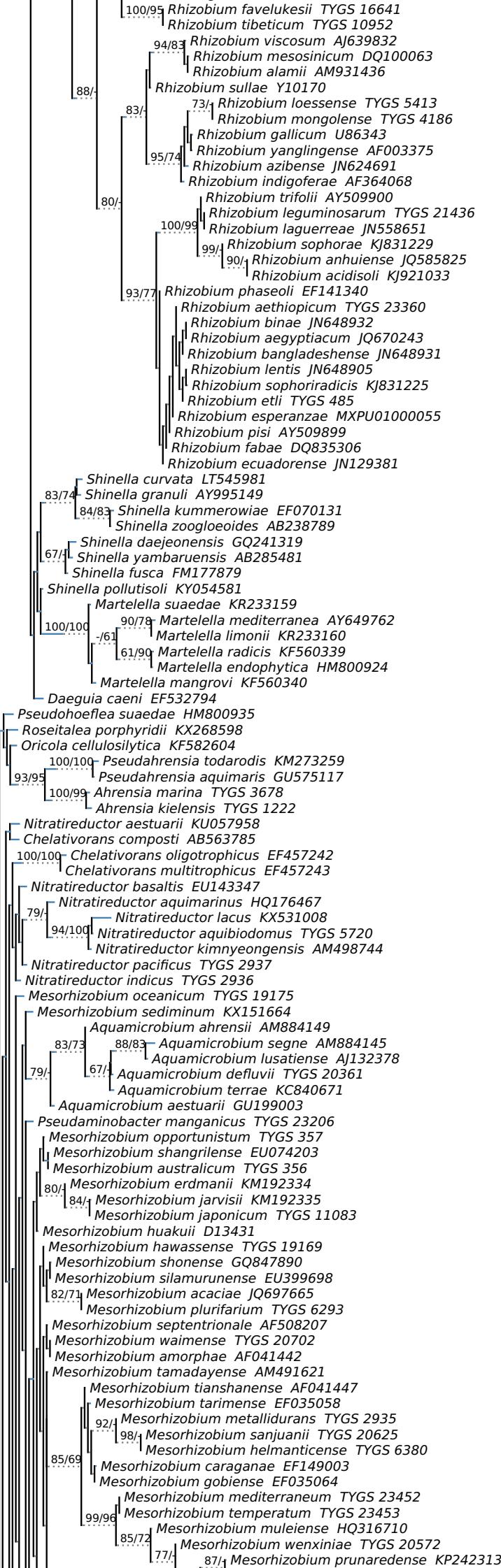


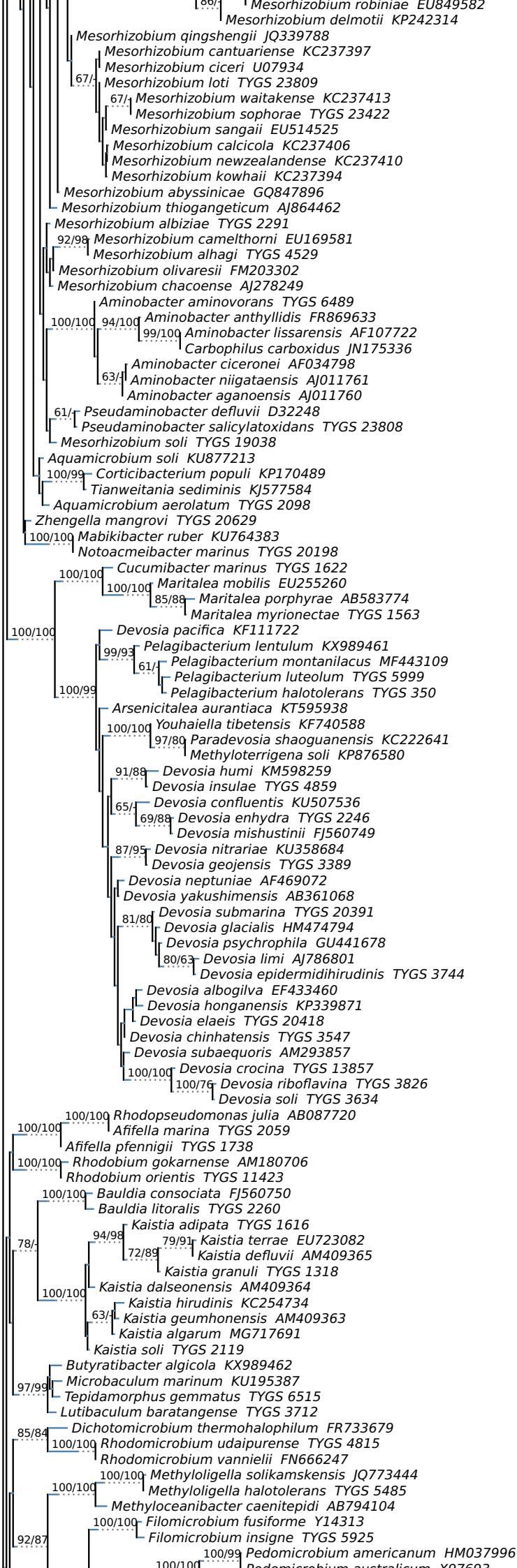


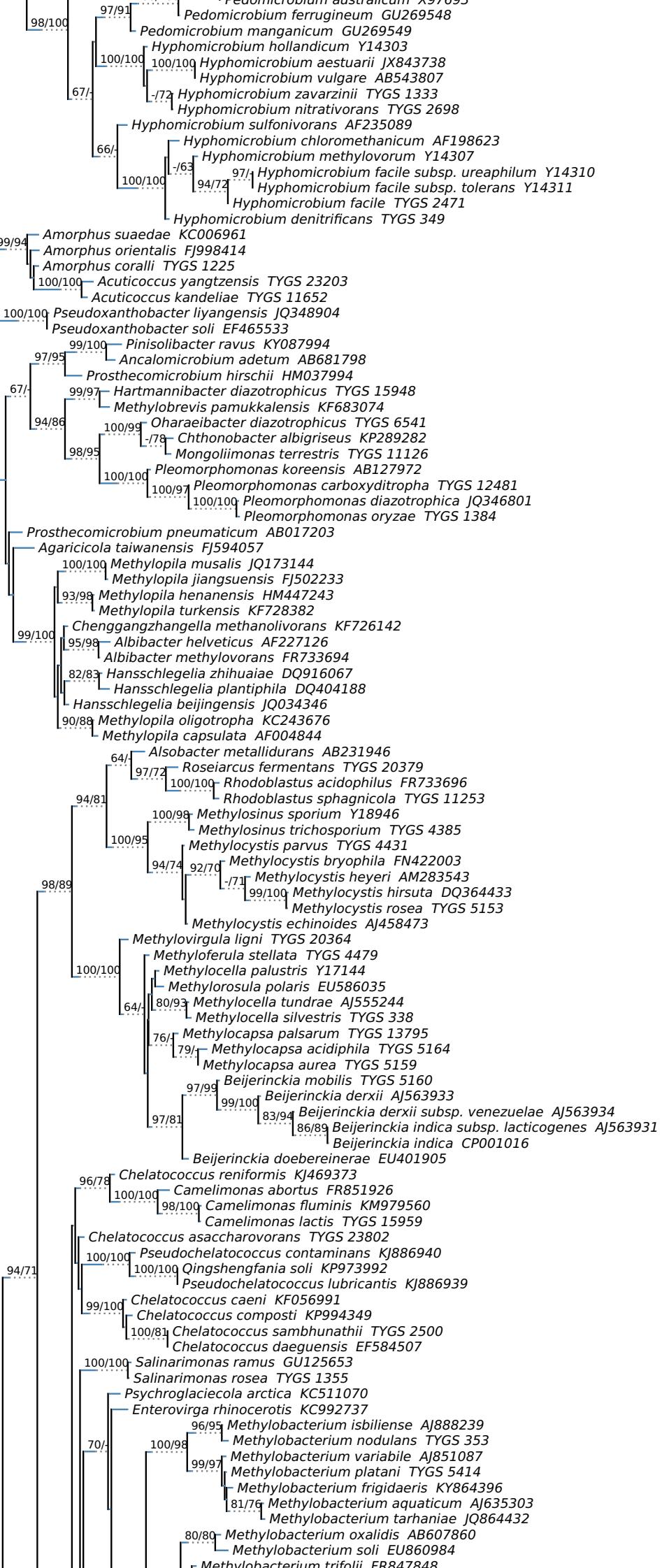












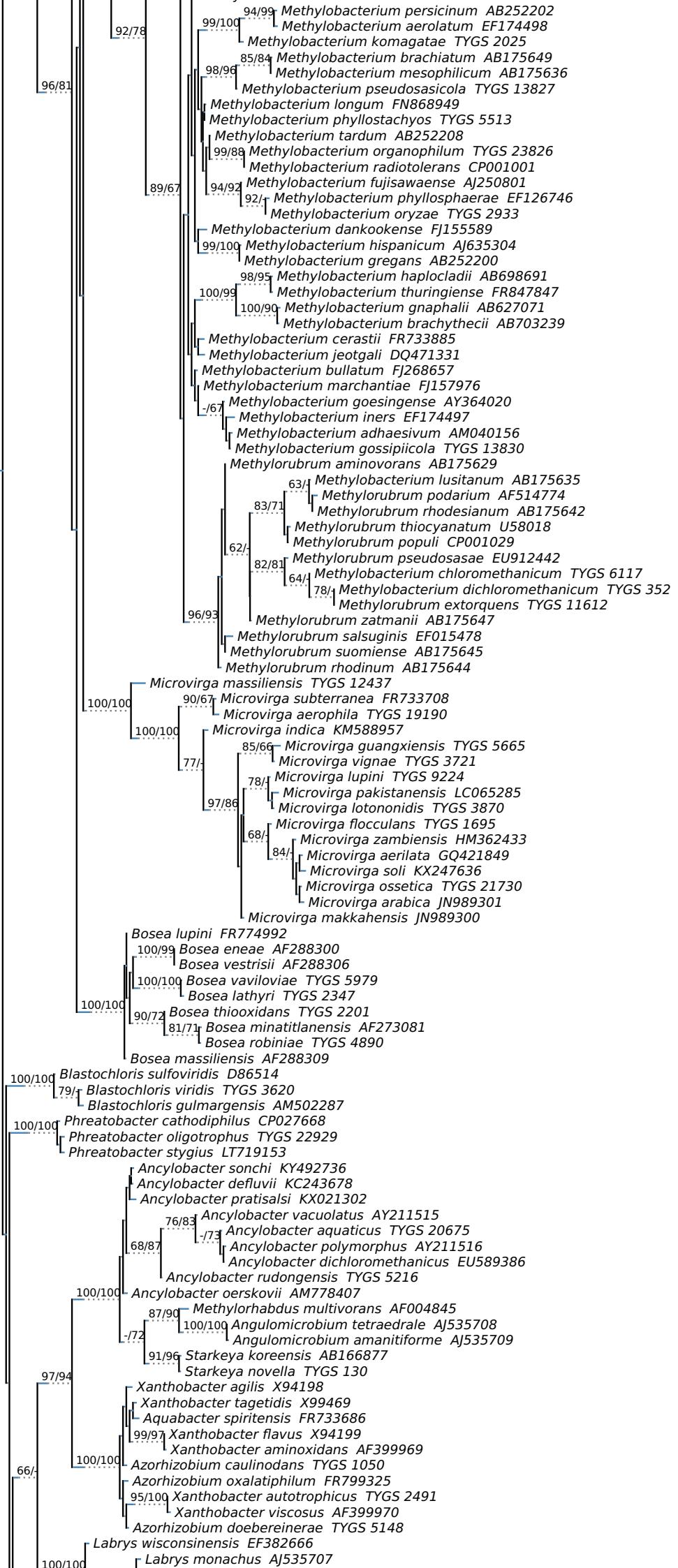
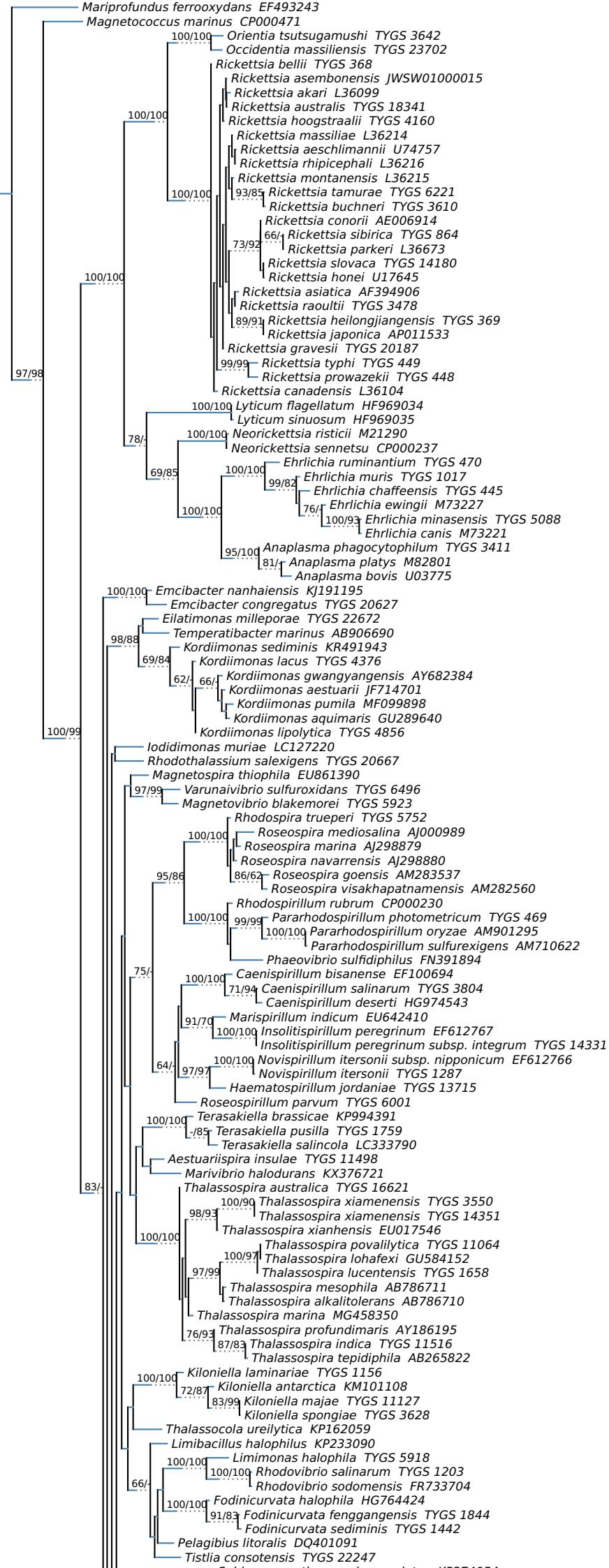
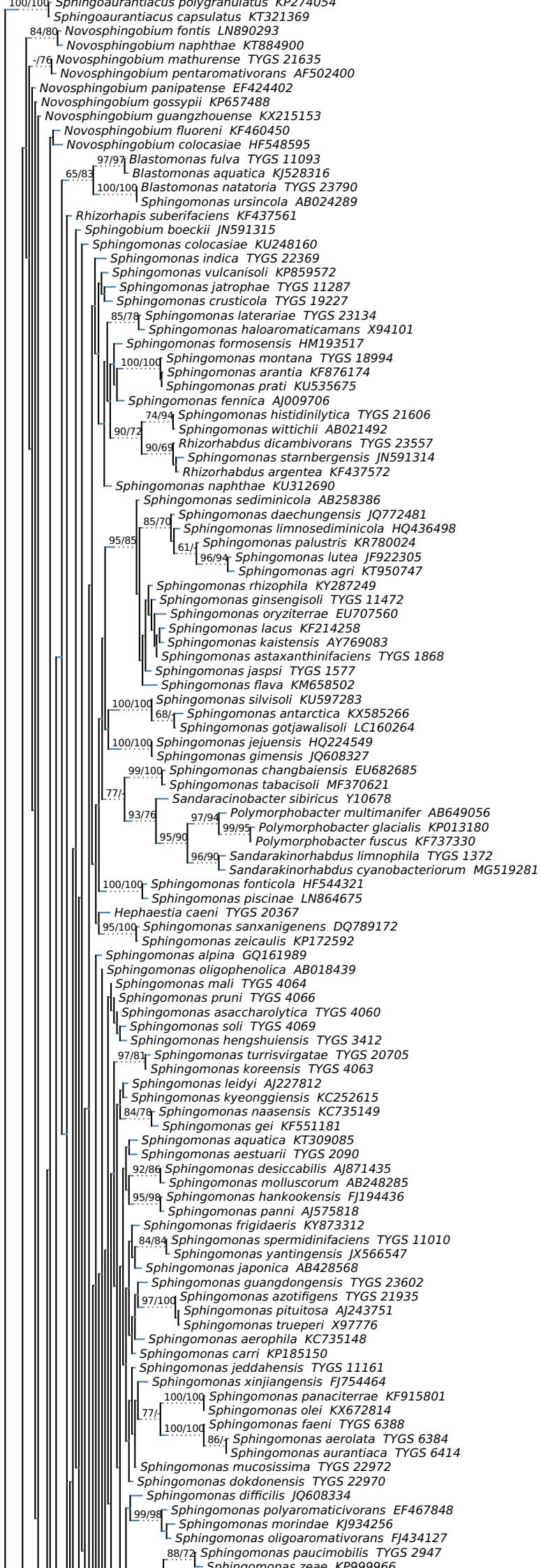


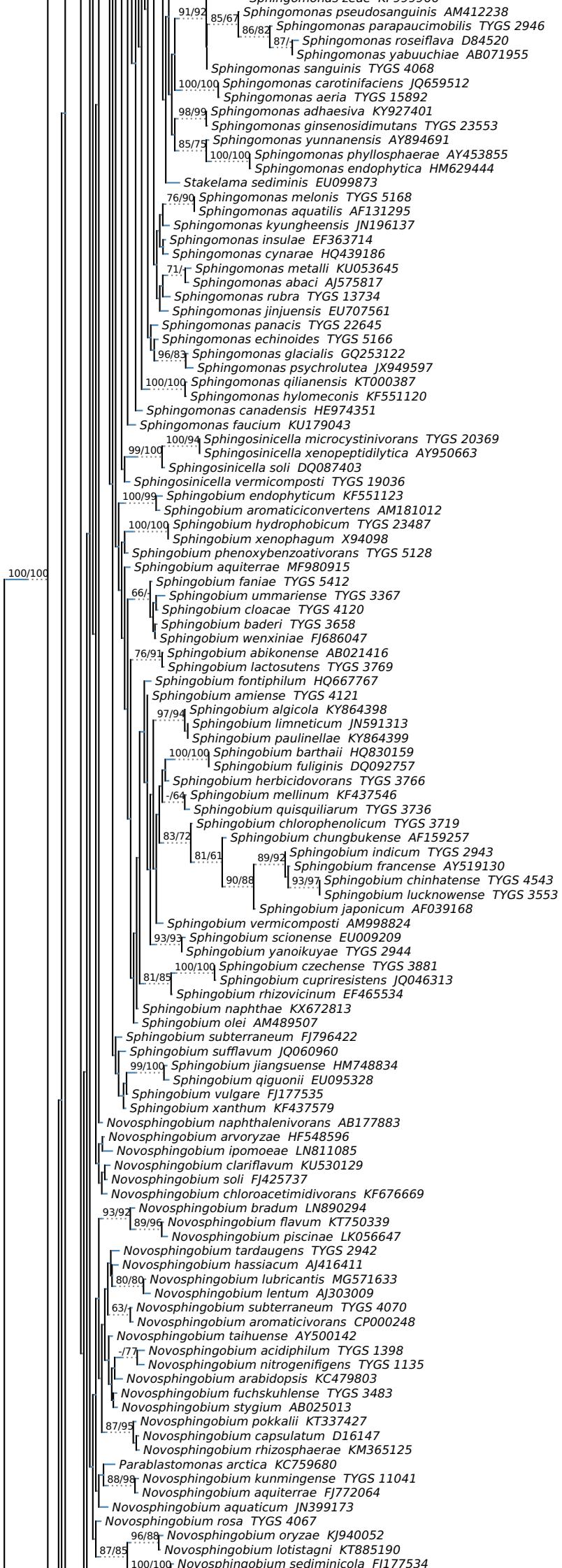


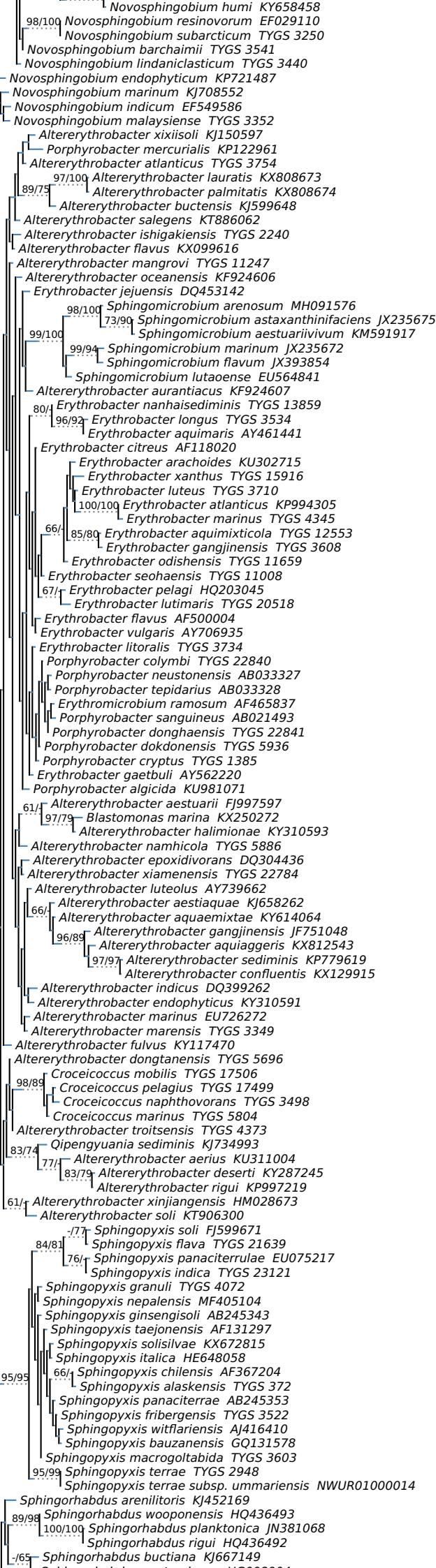


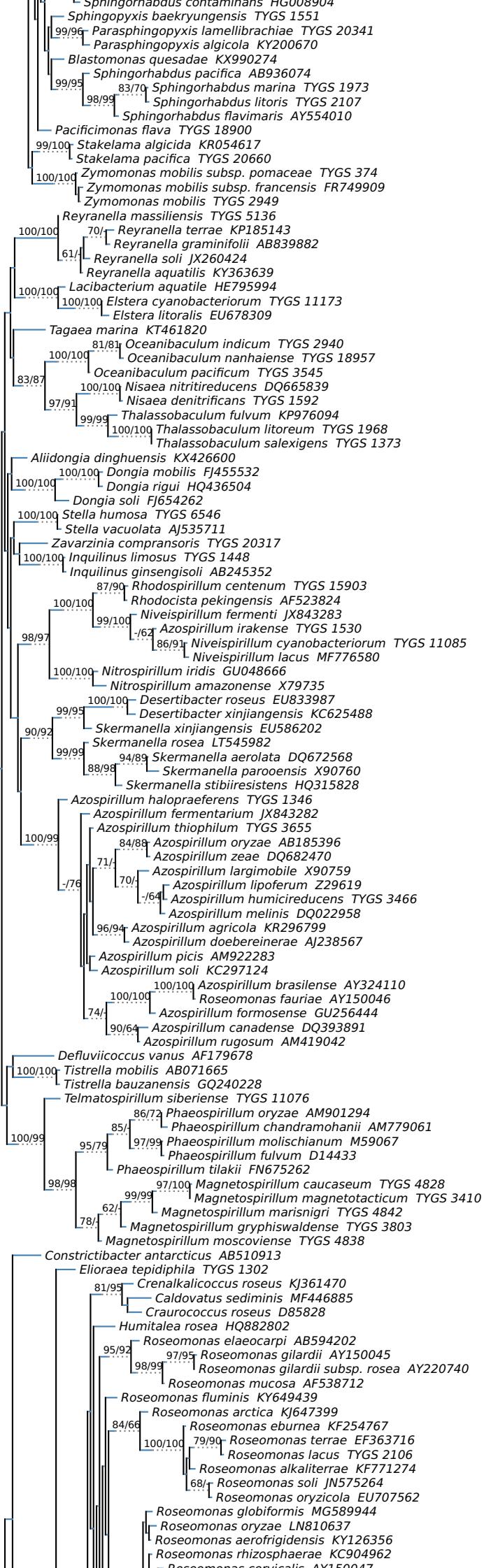
Figure 2: Backbone-constrained comprehensive 16S rRNA gene ML tree (CCT) of *Alphaproteobacteria* inferred under the GTR+CAT model. The branches are scaled in terms of the expected number of substitutions per site. The numbers above the branches are support values when larger than 60% from ML (left) and MP (right) bootstrapping. Dotted parts of branches are filled in to allow proper placement of bootstrap values and are not part of the actual branch length. Numbers preceeded by the term 'TYGS' in labels refer to the genome IDs as found in Supplementary Table S1 (first sheet). Each tip label ends with the family of the respective taxon.



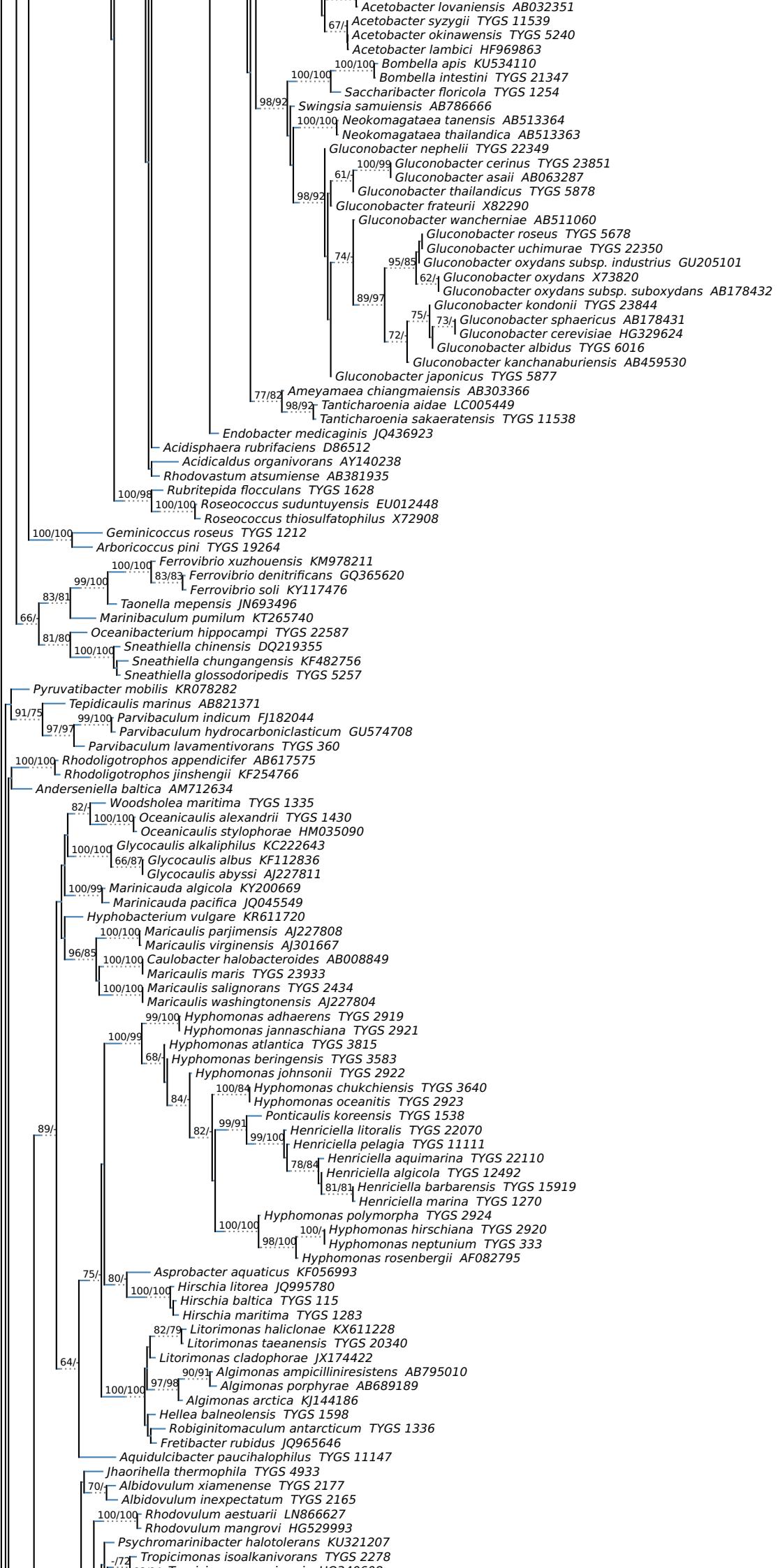


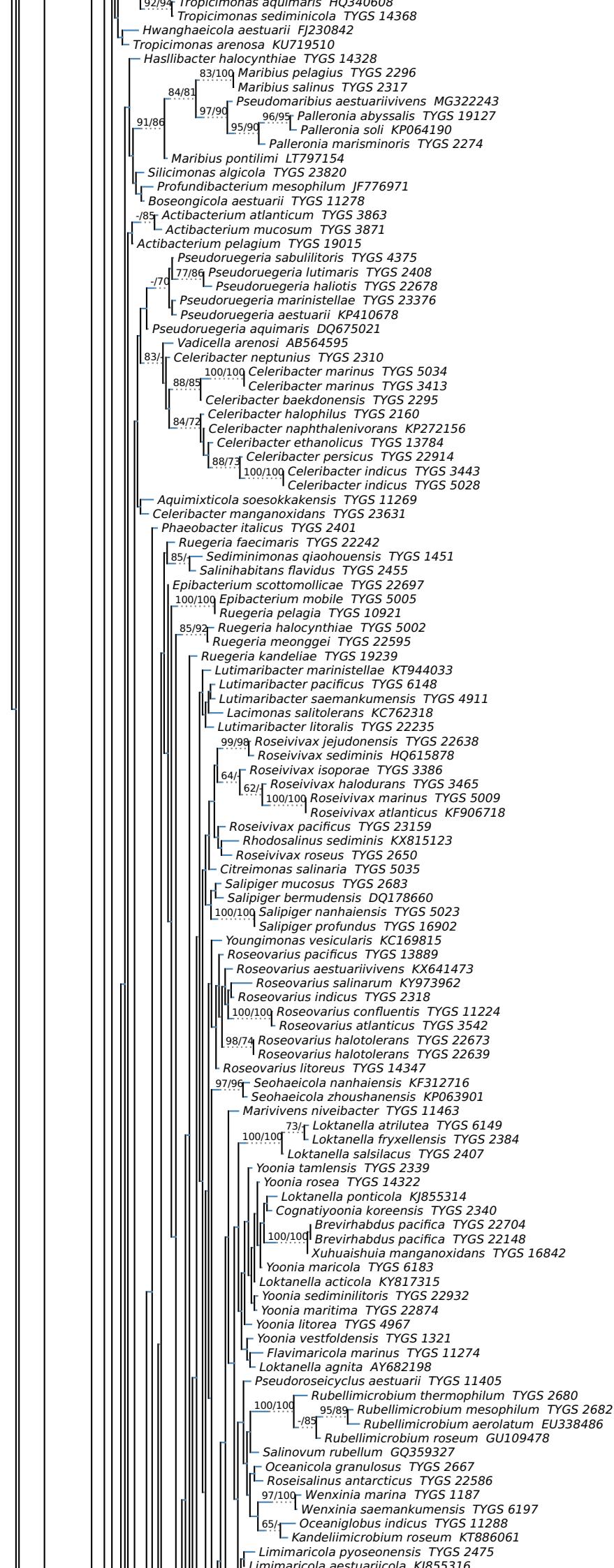


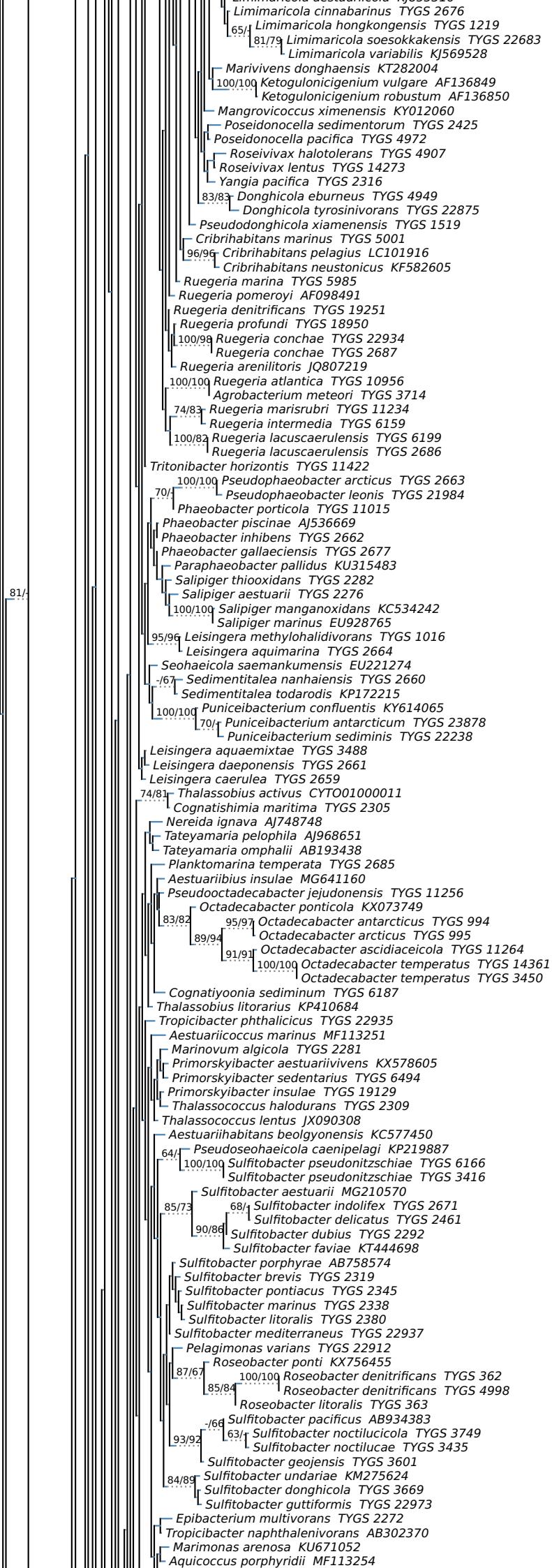


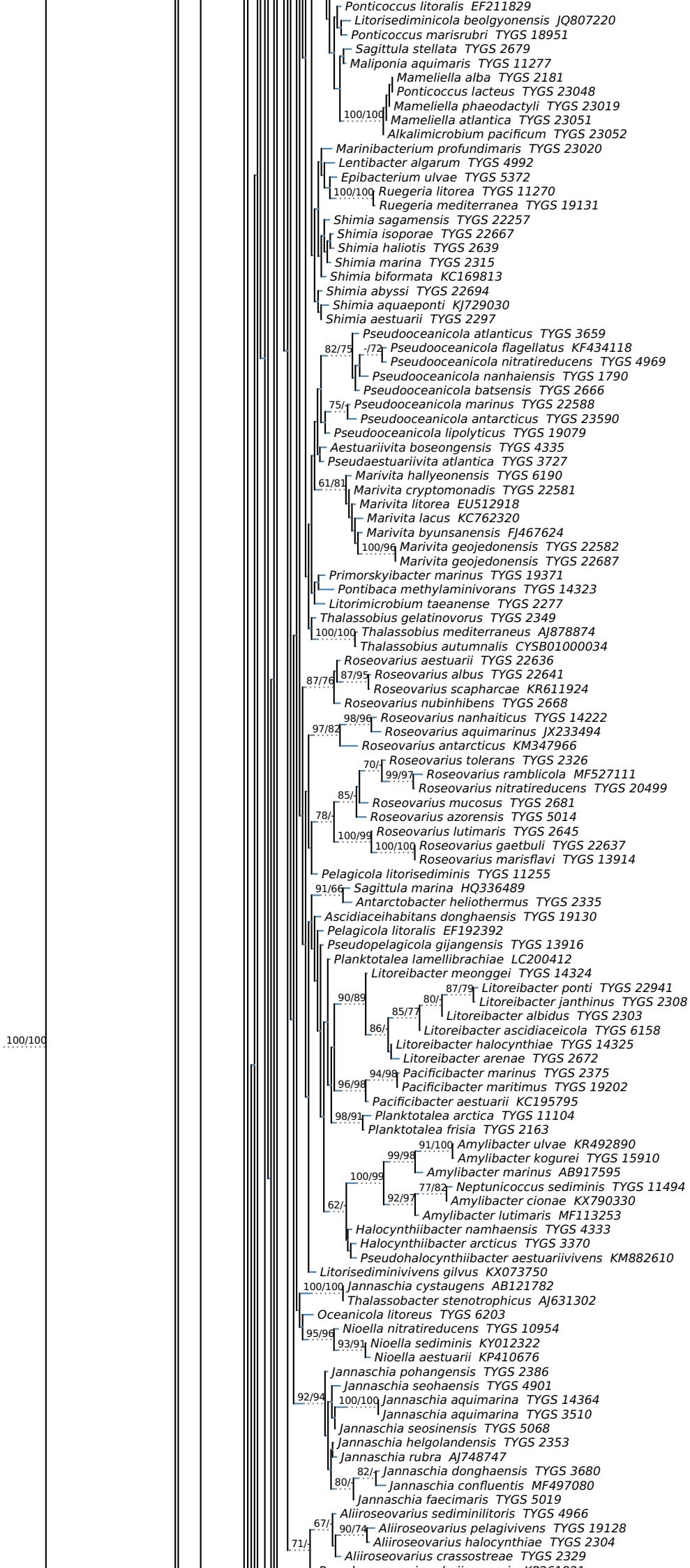


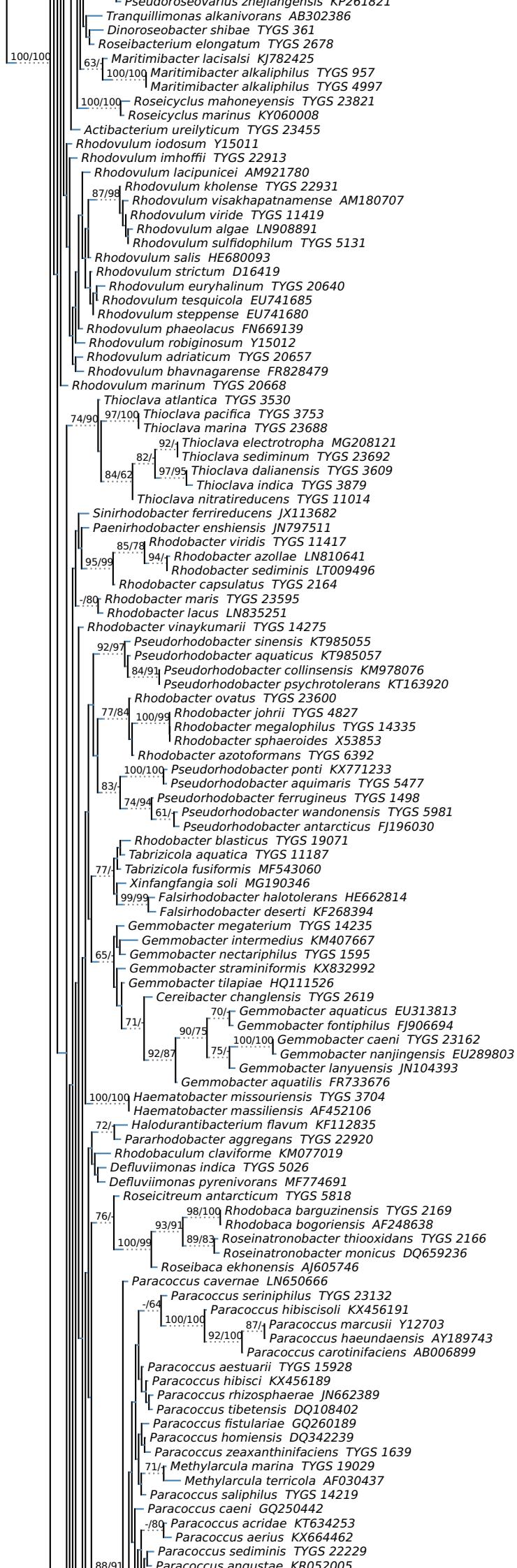


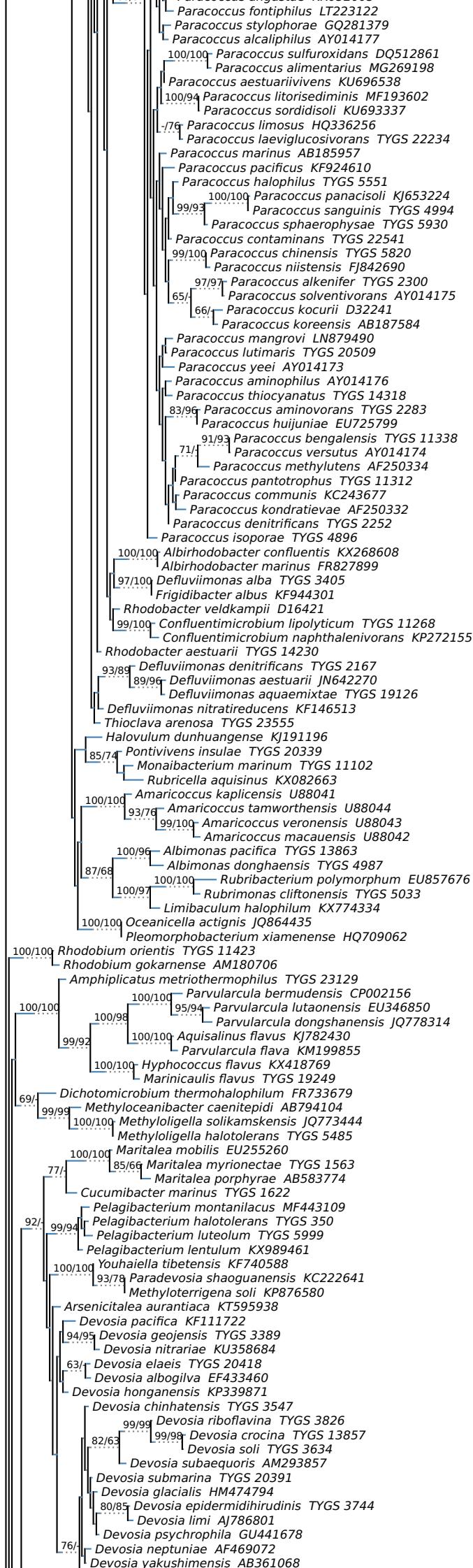


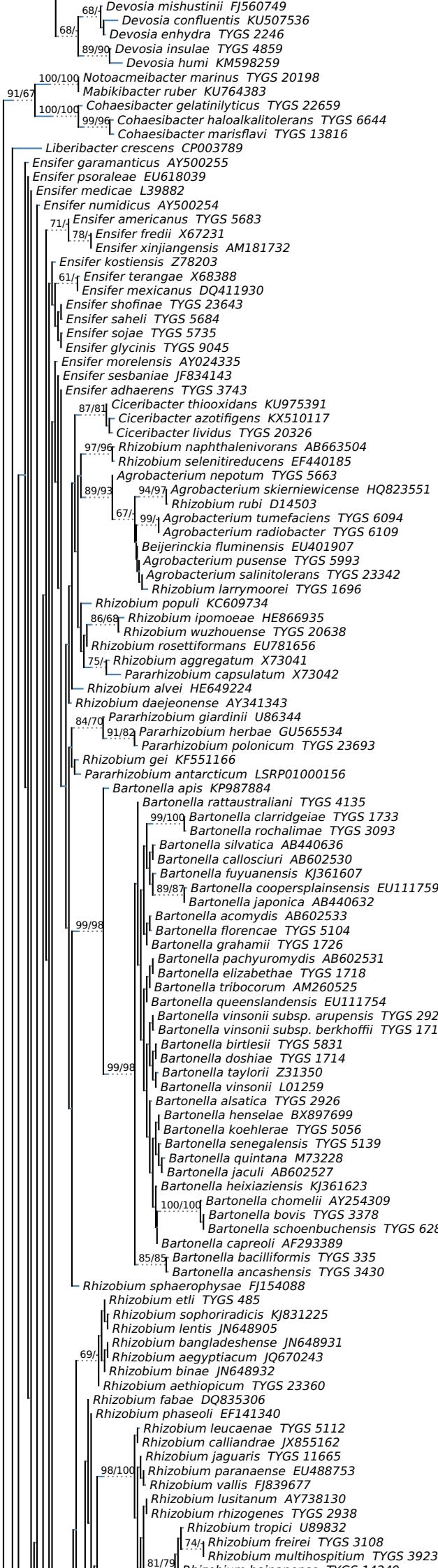


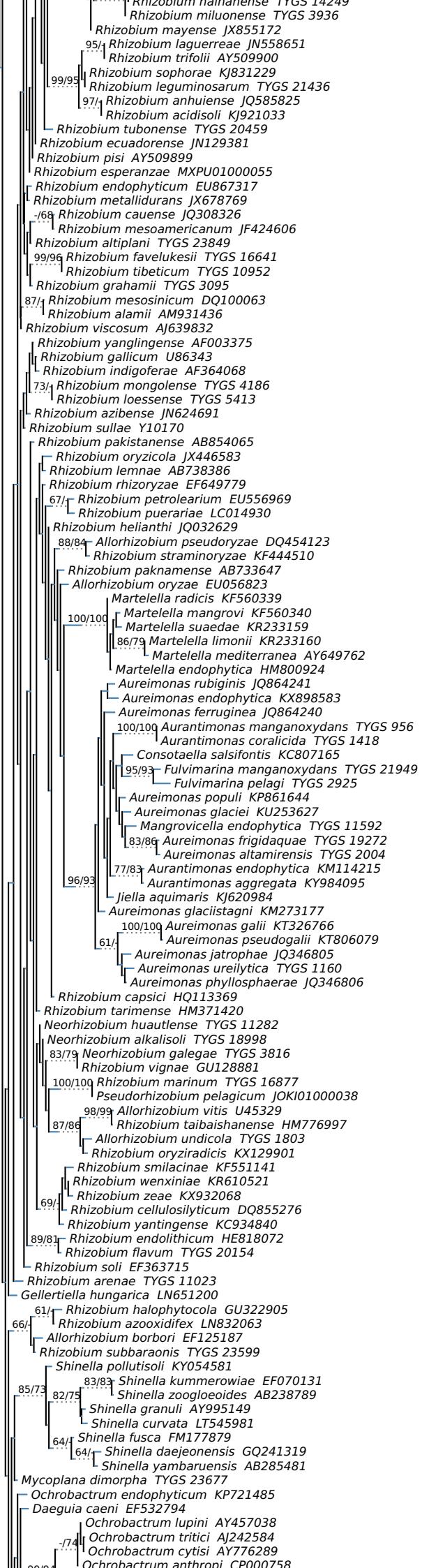


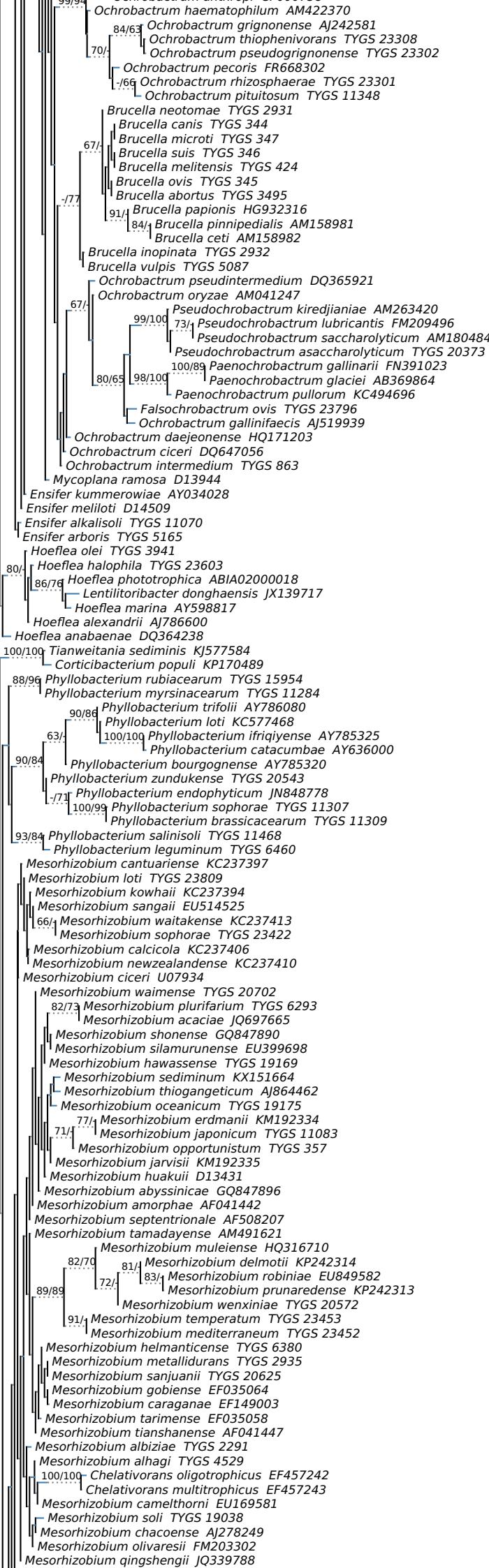


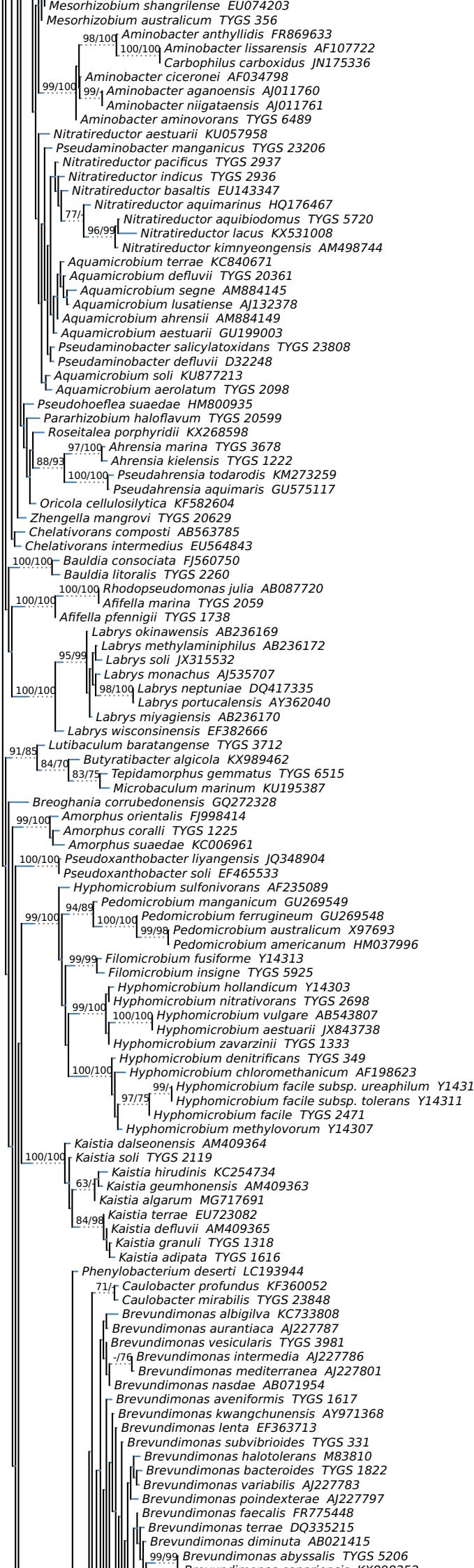


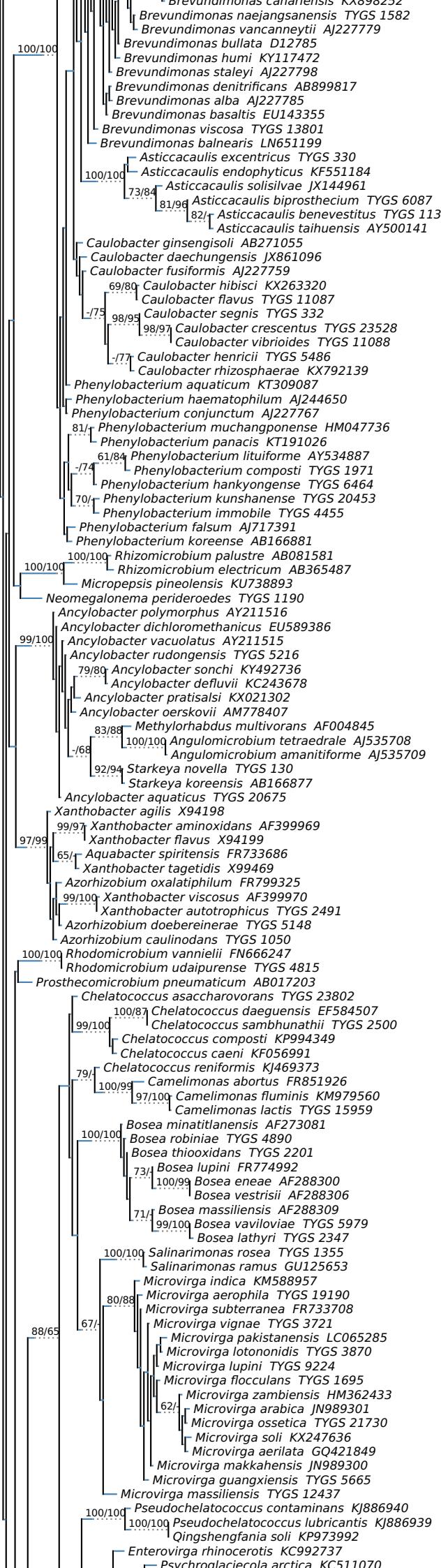


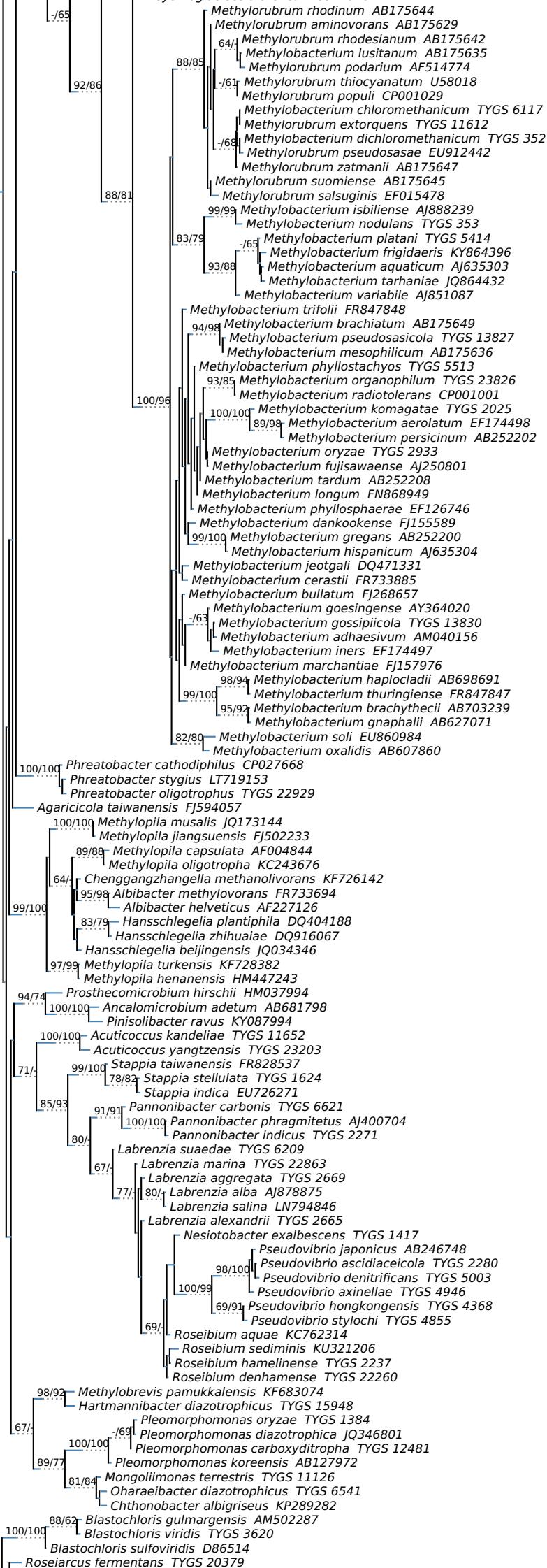
















0.25

Figure 3: Unconstrained comprehensive 16S rRNA gene ML tree (UCT) of *Alphaproteobacteria* inferred under the GTR+CAT model. The branches are scaled in terms of the expected number of substitutions per site. The numbers above the branches are support values when larger than 60% from ML (left) and MP (right) bootstrapping. Dotted parts of branches are filled in to allow proper placement of bootstrap values and are not part of the actual branch length. Numbers preceeded by the term 'TYGS' in labels refer to the genome IDs as found in Supplementary Table S1 (first sheet). Each tip label ends with the family of the respective taxon.

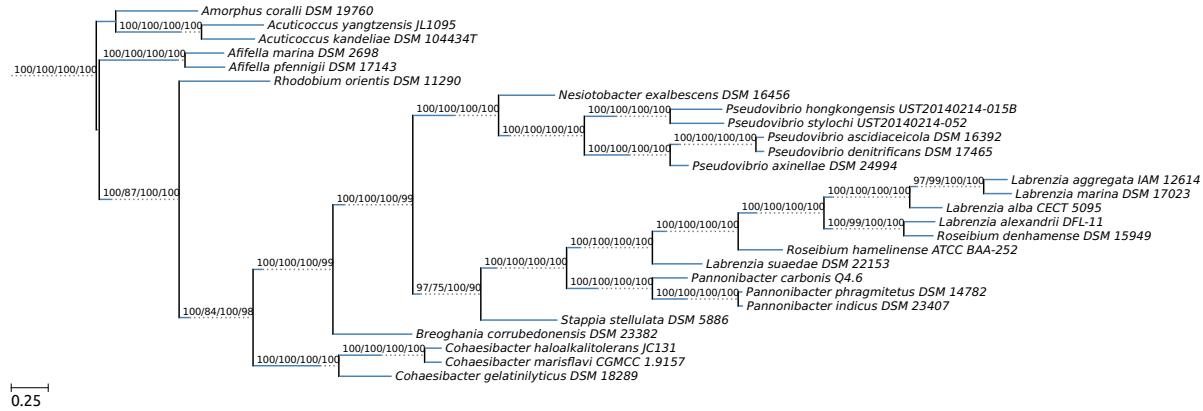


Figure 4: Phylogenetic tree inferred with RAxML from the *Labrenzia* supermatrix including single-copy core genes. Branches are scaled in terms of the expected number of changes per site. The first two numbers above branches (left to right) are partition bootstrap support values from (i) RAxML (ML) analysis and (ii) TNT (MP) analysis of the supermatrix that included single-copy genes that occurred in all of the genomes. The last two numbers above branches are partition bootstrap support values from (iii) RAxML (ML) analysis and (iv) TNT (MP) analysis of the supermatrix that included single-copy genes that occurred in at least four of the genomes.

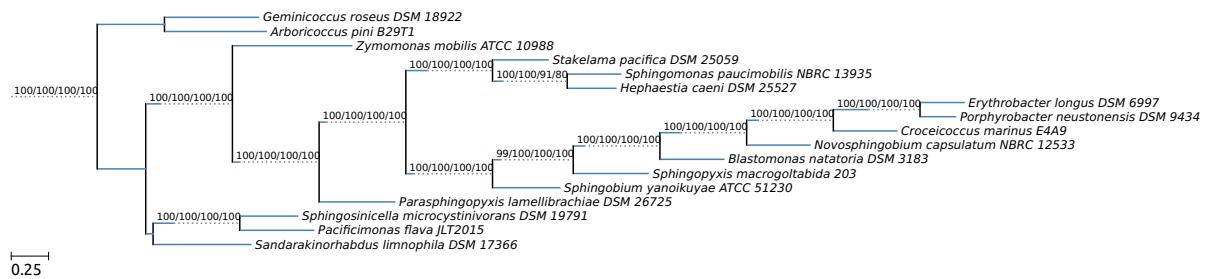


Figure 5: Phylogenetic ML tree inferred with RAxML from the *Sphingomonadales* supermatrix including single-copy core genes. Branches are scaled in terms of the expected number of changes per site. The first two numbers above branches (left to right) are partition bootstrap support values from (i) RAxML (ML) analysis and (ii) TNT (MP) analysis of the supermatrix that included single-copy genes that occurred in all of the genomes. The last two numbers above branches are partition bootstrap support values from (iii) RAxML (ML) analysis and (iv) TNT (MP) analysis of the supermatrix that included single-copy genes that occurred in at least four of the genomes.