# Supplementary Material

## A proposal for a standardized bacterial taxonomy based on genome phylogeny

Donovan H. Parks[1], Maria Chuvochina[1], David W. Waite[1], Christian Rinke[1], Adam Skarshewski[1], Pierre-Alain Chaumeil[1], Philip Hugenholtz[1]

1. Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, QLD 4072, Australia

**The Supplementary material includes:**

Methods and Materials
Supp. Tables 1 to 9
Supp. Figures 1 to 5

## Supplementary Tables

All tables are provided in Excel format.

**Supp. Table 1**. Robustness of GTDB taxonomy under varying marker sets, subsets of taxa, and evolutionary models. Bootstrap support and relative evolutionary divergence (RED) values are given for trees inferred from 120 concatenated proteins (bac120), 16 concatenated ribosomal proteins (rp1), and the 16S rRNA gene (16S) using the WAG substitution model. Results are also given for a tree inferred from bac120 under the LG model. F-measure, precision, and recall statistics indicating the fit of GTDB taxa on the rp1, 16S, and LG trees are provided. Trees were also inferred from subsets of the 120 proteins (bac120 gene subsample), under taxon resampling with one genome per genus (bac120 genus subsample), and for each gene within the 120 protein set (bac120 gene trees). For these cases, the percentage of trees resulting in a GTDB taxon being monophyletic, operationally monophyletic, or polyphyletic is indicated.

**Supp. Table 2.** 16S rRNA-based taxa names adopted in the GTDB taxonomy and their associated rank and number of circumscribed genomes.

**Supp. Table 3.** Correspondence between standardly named NCBI and GTDB taxa ordered by degree of polyphyly. Percentage of genomes assigned to each GTDB taxon for a given NCBI taxon is shown in parentheses. Taxa for each rank from phylum to genus are displayed in separate sheets.

**Supp. Table 4.** Genomes with conflicting or unresolved taxonomic assignments when applying the GTDB taxonomy to a tree inferred from the concatenation of 16 ribosomal proteins (rp1).

**Supp. Table 5.** Genomes with conflicting or unresolved taxonomic assignments when applying the GTDB taxonomy to a tree inferred from the 16S rRNA gene.

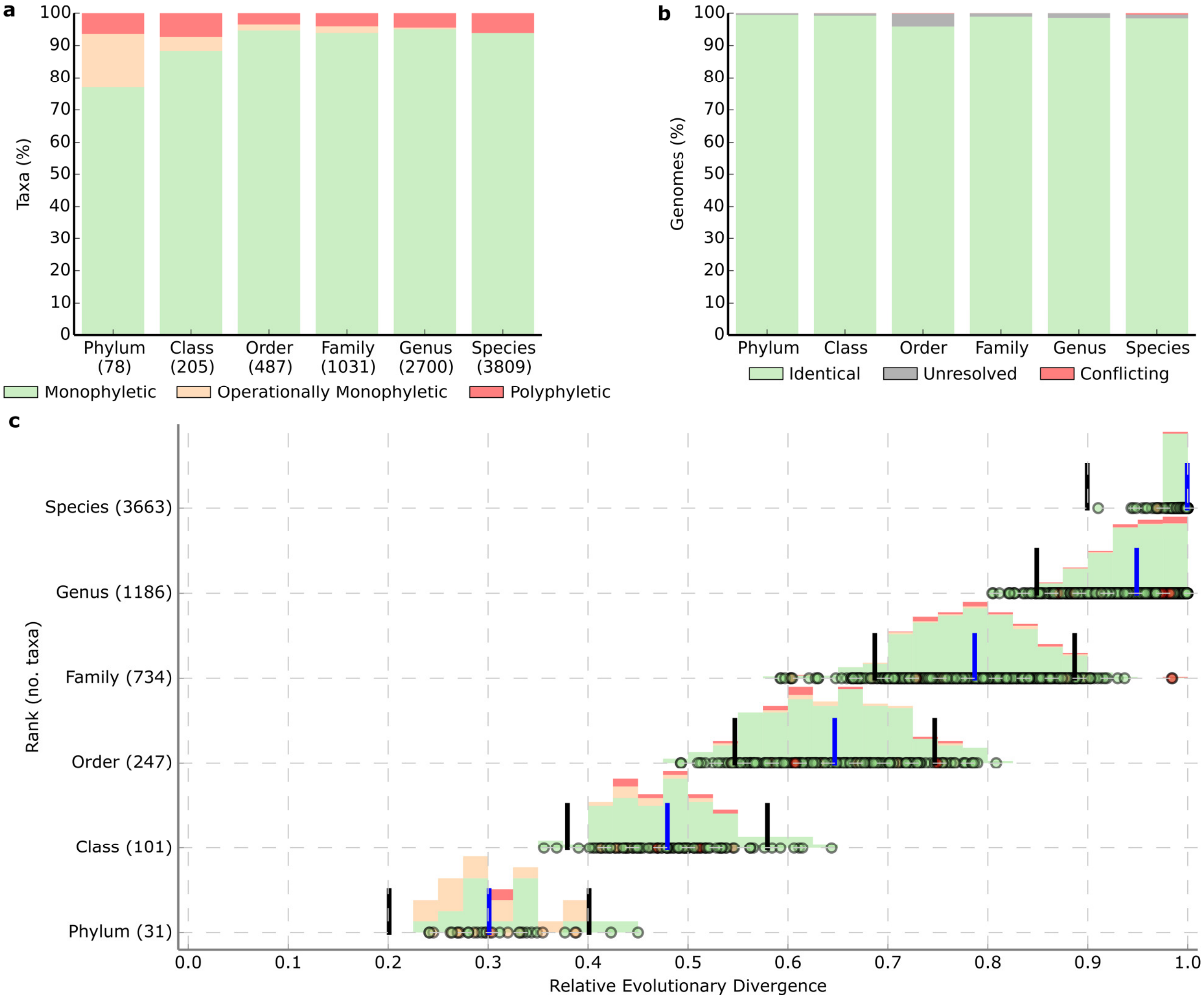
**Supp. Table 6.** NCBI taxa that have been 'retired' in the GTDB taxonomy and brief explanations for their retirement.

**Supp. Table 7.** Comparison of NCBI and GTDB genus and species classifications to those proposed by Beaz-Hidalgo et al. (2015), Kook et al. (2017), and Bobay & Ochman (2017).

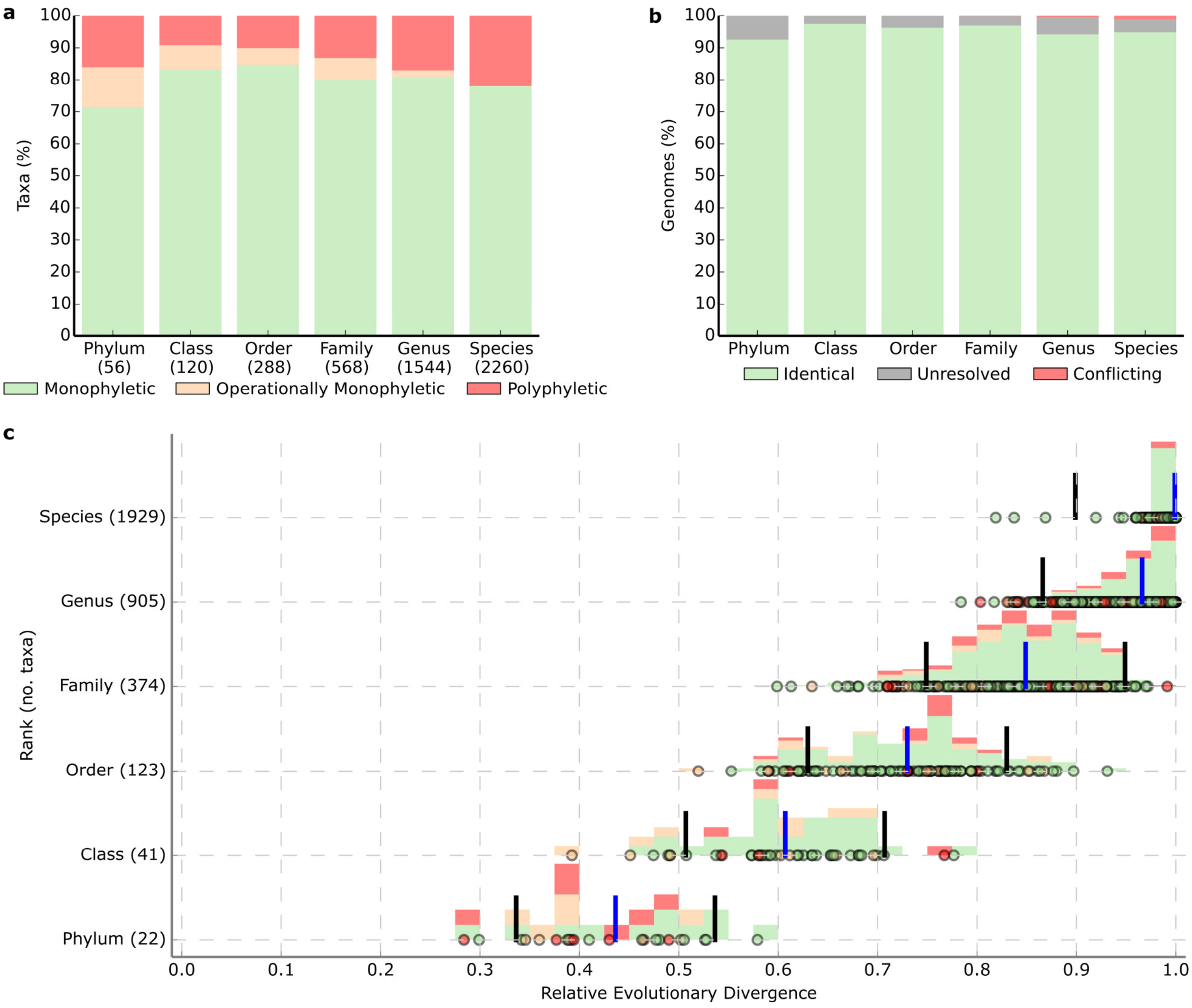
**Supp. Table 8.** Comparison of clostridia classifications proposed by Yutin & Galperin (2013) to the GTDB taxonomy.

**Supp. Table 9**. Draft genomes with 16S rRNA genes that do not meet the selection criteria for inclusion in the 16S rRNA tree (*see Methods*).

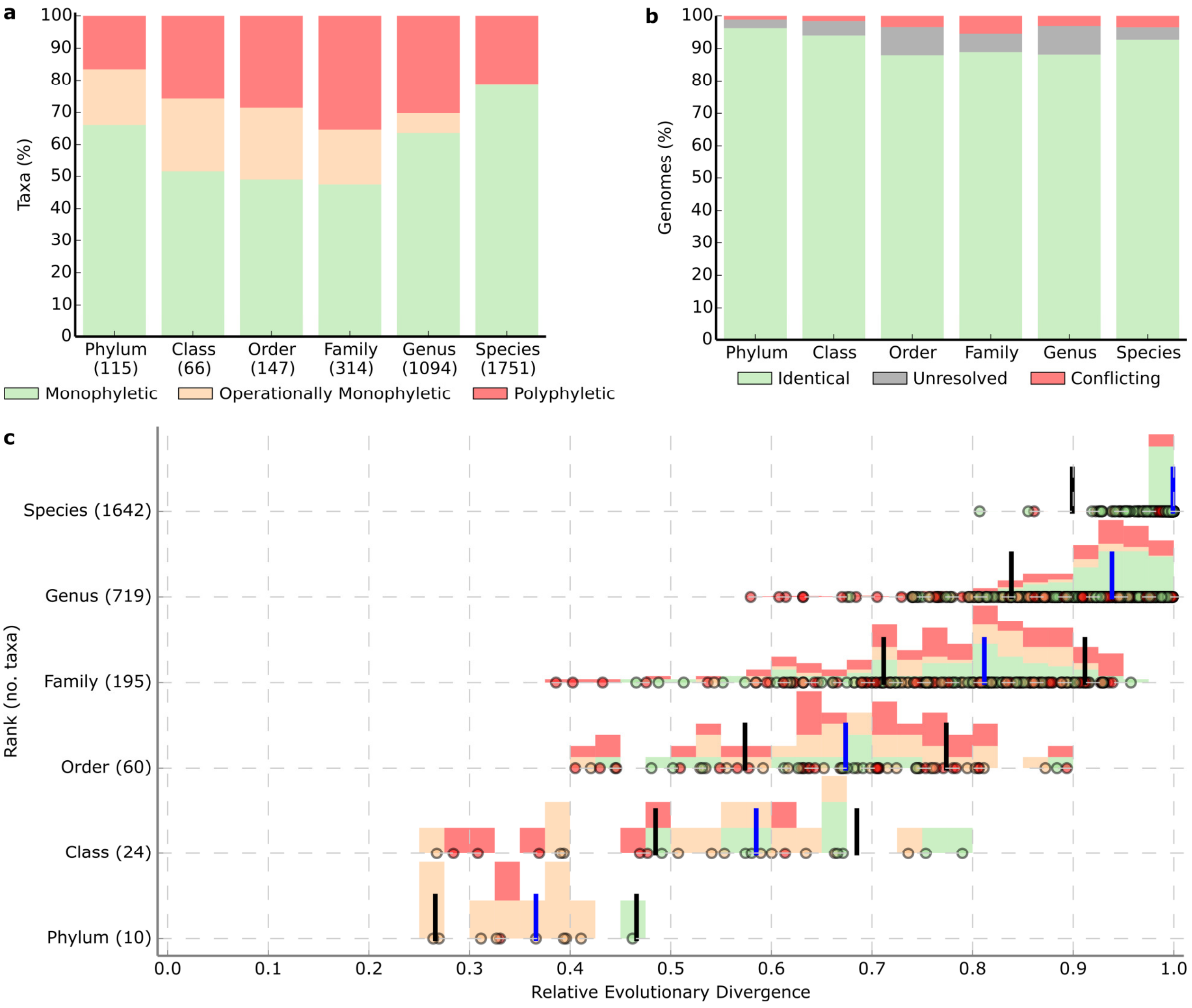
## Supplementary Figures



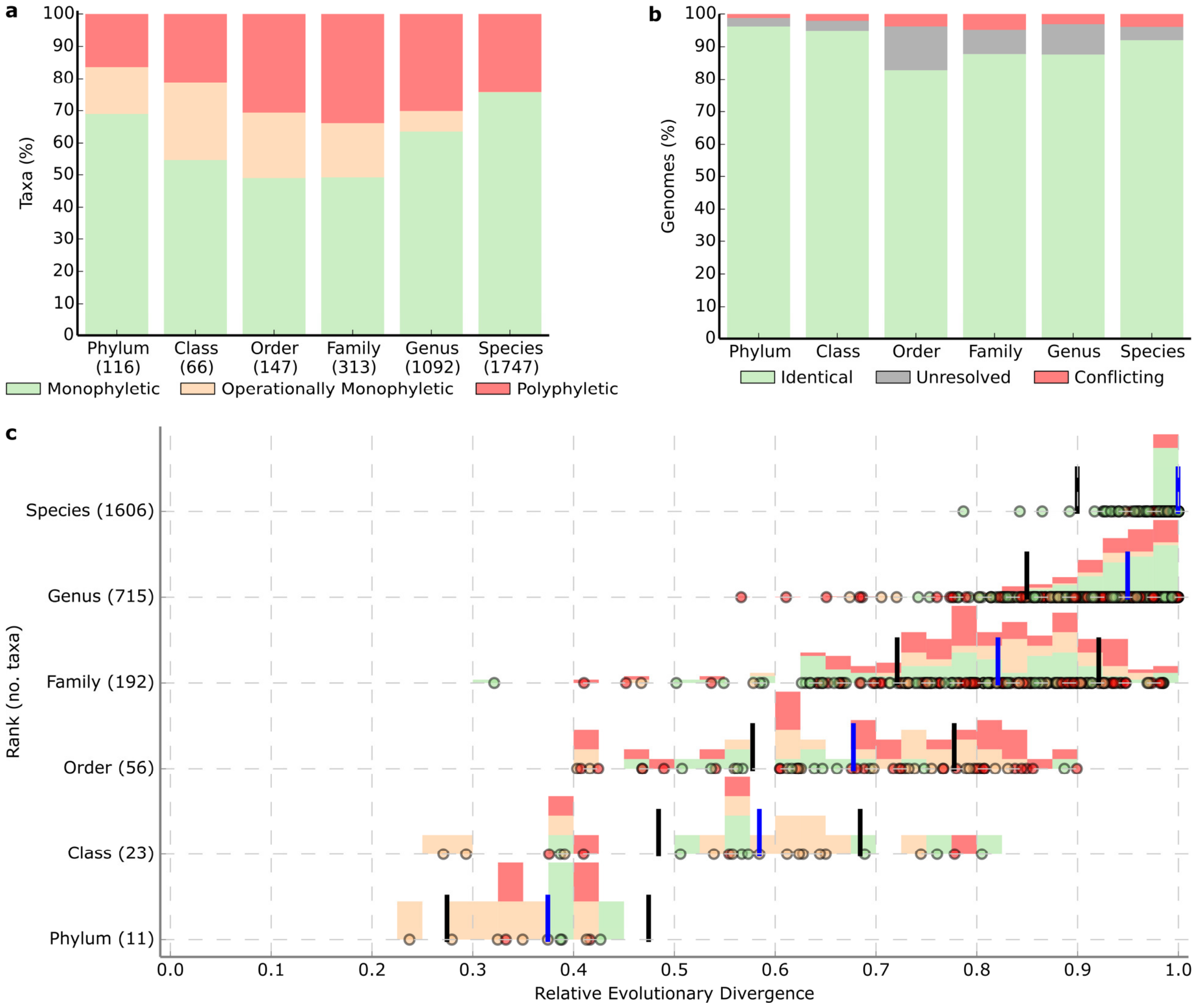
**Supp. Figure 1. Congruence of the GTDB taxonomy on a tree inferred from the concatenation of 16 ribosomal proteins (rp1).** (**a**) Percentage of GTDB taxa at each rank which are monophyletic, operationally monophyletic, or polyphyletic within the rp1 tree. Results were calculated over all taxa comprised of >1 genomes and the number of taxa considered at each rank is shown in parentheses. (**b**) Percentage of the 20,699 genomes within the rp1 tree with identical, unresolved, or conflicting taxonomic assignments at each rank relative to the GTDB taxonomy when taxonomy is assigned based on their placement in the rp1 tree. (**c**) RED of taxa with ≥2 immediate subordinate taxa in the rp1 tree, with the same coloring as used in panel a. The number of taxa plotted at each rank is given in parentheses.



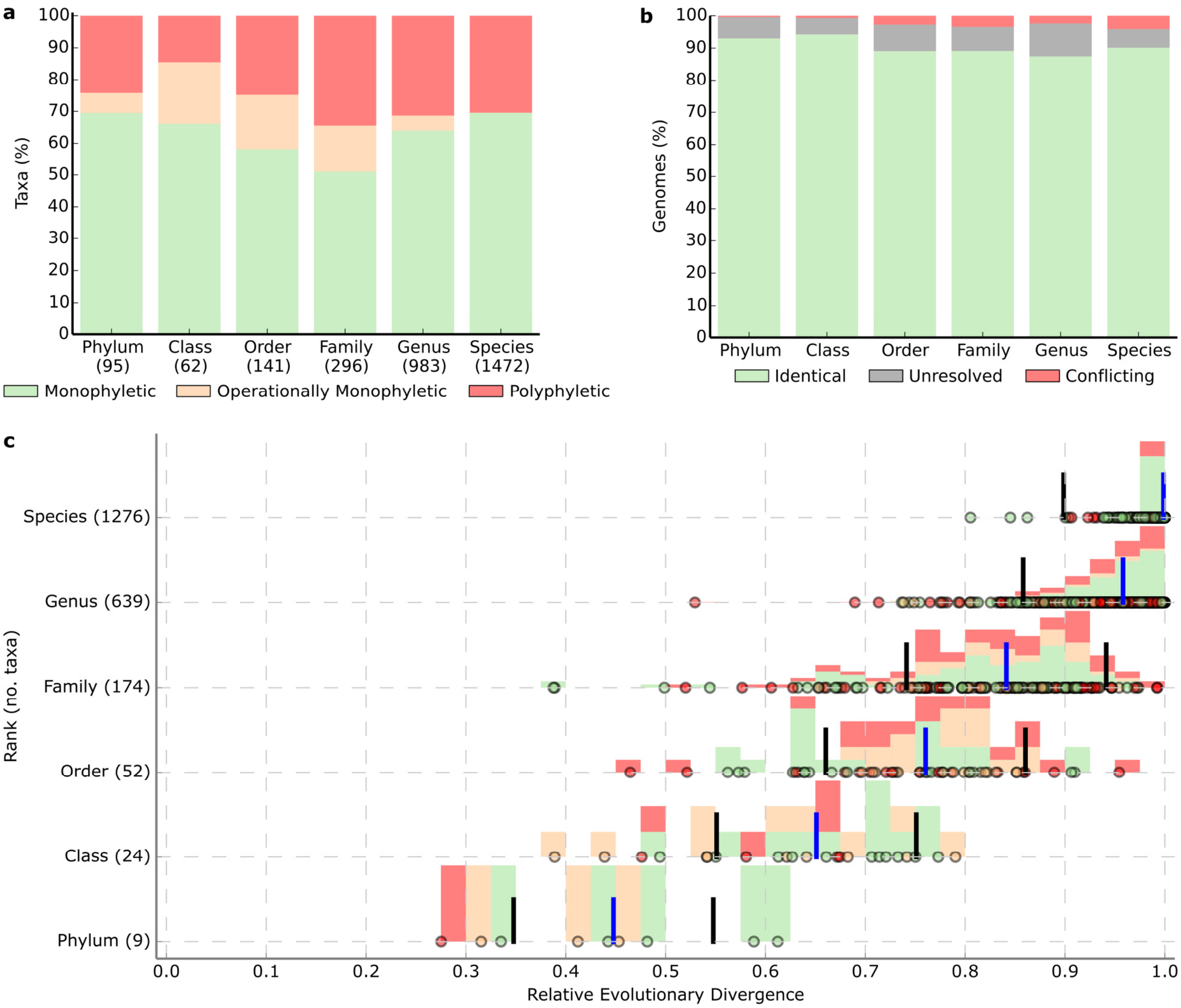
**Supp. Figure 2**. **Congruence of the GTDB taxonomy on a tree inferred from the 16S rRNA gene**. (**a**) Percentage of GTDB taxa at each rank which are monophyletic, operationally monophyletic, or polyphyletic within the 16S rRNA gene tree. Results were calculated over all taxa comprised of >1 genomes and the number of taxa considered at each rank is shown in parentheses. (**b**) Percentage of the 11,243 genomes within the 16S rRNA gene tree with identical, unresolved, or conflicting taxonomic assignments at each rank relative to the GTDB taxonomy when taxonomy is assigned based on their placement in the 16S rRNA gene tree. (**c**) RED of taxa with ≥2 immediate subordinate taxa in the 16S rRNA gene tree, with the same coloring as used in panel a. The number of taxa plotted at each rank is given in parentheses.



**Supp. Figure 3. Congruence of the NCBI taxonomy on a tree inferred from the concatenation of 120 proteins (bac120)**. (**a**) Percentage of NCBI taxa at each rank which are monophyletic, operationally monophyletic, or polyphyletic within the bac120 tree. Results were calculated over all taxa comprised of >1 genomes and the number of taxa considered at each rank is shown in parentheses. (**b**) Percentage of the 16,248 RefSeq/GenBank genomes within the bac120 tree with identical, unresolved, or conflicting taxonomic assignments at each rank relative to the NCBI taxonomy when taxonomy is assigned based on their placement in the protein tree. (**c**) RED of taxa with ≥2 immediate subordinate taxa in the bac120 tree, with the same coloring as used in panel a. The number of taxa plotted at each rank is given in parentheses (note that this is a reproduction of **Fig. 2a**).



**Supp. Figure 4**. **Congruence of the NCBI taxonomy on a tree inferred from the concatenation of 16 ribosomal proteins (rp1)**. (**a**) Percentage of NCBI taxa at each rank which are monophyletic, operationally monophyletic, or polyphyletic within the rp1 tree. Results were calculated over all taxa comprised of >1 genomes and the number of taxa considered at each rank is shown in parentheses. (**b**) Percentage of the 16,306 RefSeq/GenBank genomes within the rp1 tree with identical, unresolved, or conflicting taxonomic assignments at each rank relative to the NCBI taxonomy when taxonomy is assigned based on their placement in the rp1 tree. (**c**) RED of taxa with ≥2 immediate subordinate taxa in the rp1 tree, with the same coloring as used in panel a. The number of taxa plotted at each rank is given in parentheses.



**Supp. Figure 5**. **Congruence of the NCBI taxonomy on a tree inferred from the 16S rRNA gene**. (**a**) Percentage of NCBI taxa at each rank which are monophyletic, operationally monophyletic, or polyphyletic within the 16S rRNA gene tree. Results were calculated over all taxa comprised of >1 genomes and the number of taxa considered at each rank is shown in parentheses. (**b**) Percentage of the 11,147 RefSeq/GenBank genomes within the 16S rRNA gene tree with identical, unresolved, or conflicting taxonomic assignments at each rank relative to the NCBI taxonomy when taxonomy is assigned based on their placement in the 16S rRNA gene tree. (**c**) RED of taxa with ≥2 immediate subordinate taxa in the 16S rRNA gene tree, with the same coloring as used in panel a. The number of taxa plotted at each rank is given in parentheses.

## References

Choi EJ et al. 2013. *Mooreia alkaloidigena* gen. nov., sp. nov. and *Catalinimonas alkaloidigena* gen. nov., sp. nov., alkaloid-producing marine bacteria in the proposed families *Mooreiaceae* fam. nov. and *Catalimonadaceae* fam. nov. in the phylum *Bacteroidetes*. *Int J Syst Evol Microbiol* **63**: 1219-1228.

Joseph SJ et al. 2003. Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl Environ Microbiol* **69**: 7210-7215.

Kublanov IV, et al. 2017. Genomic analysis of *Caldithrix abyssi*, the thermophilic anaerobic bacterium of the novel bacterial phylum *Calditrichaeota*. *Front Microbiol* **8**: 195.

Moreira D, et al. 2017. Description of *Gloeomargarita lithophora* gen. nov., sp. nov., a thylakoid-bearing, basal-branching cyanobacterium with intracellular carbonates, and proposal for Gloeomargaritales ord. nov. *Int J Syst Evol Microbiol* **67**: 653-658.

Naushad S et al. 2015. A phylogenomic and molecular marker based taxonomic framework for the order *Xanthomonadales*: proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales*, the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. *Antonie van Leeuwenhoek* **107**: 467-485.