

Supplemental Material to:

Paweł Mackiewicz, Andrzej Bodył, Krzysztof Moszczyński

The case of horizontal gene transfer from bacteria to the peculiar dinoflagellate plastid genome

Mobile Genetic Elements 2013; 3(4)

<http://dx.doi.org/10.4161/mge.25845>

<http://www.landesbioscience.com/journals/mge/article/25845/>

Materials and methods

Collection of sequences and preparation of alignments

To gather all potential homologs to Ycf16 and Ycf24 encoded on the *Ceratium horridum* AF490364 minicircle as well as the sequences of Rpl28, Rpl33, and the potential product of the unannotated open reading frame (FtsY) encoded on the *Pyrocystis lunula* AF490367 minicircle, we carried out comprehensive PSI-BLAST or BLAST searches of non-redundant protein and expressed sequence tag databases in NCBI GenBank (<http://blast.ncbi.nlm.nih.gov>), Dragonblast (<http://dbdata.rutgers.edu/dragon>), and the expressed sequence tag database of *Alexandrium tamarense*.¹ The identified sequences were verified by local searches of the Conserved Domain Database² for the presence of appropriate domains. Each of the resulting sets of homologs consisted of several thousand sequences. Initial alignments were performed in MAFFT using the slow but accurate algorithm L-INS-i with 1,000 cycles of iterative refinement³ and edited manually in JalView.⁴ Incomplete or fragmentary sequences were excluded from further analyses. Ultimately, more than 100 sequences in each of five protein sets were selected using T-Coffee⁵ to remove redundancy from the datasets and to include representatives from various prokaryotic and eukaryotic groups. Final alignments were obtained in T-Coffee using profile information (PSI-Coffee) and combining the output of many alignment methods (M-Coffee).

Phylogenetic analyses

Phylogenetic trees were inferred by six approaches using five programs: PhyloBayes 3.3e (ref. 6), MrBayes 3.2.1 (ref. 7), TreeFinder⁸, PhyML-Structure⁹, and morePhyML 1.14 (ref. 10) based on PhyML 3.0 (ref. 11). In the PhyloBayes analyses, we applied two substitution models for all alignment sets, LG+ Γ (5) and CAT Poisson+ Γ (5), with the number of components, weights, and profiles inferred from the data. Two independent Markov chains were run through 100,000 cycles (for the Ycf24 set) or 200,000 cycles (for the other sets) with the former model and through 1,000,000 cycles with the latter model. A posterior consensus was calculated from the last 10,000-500,000 trees from each chain after obtaining convergence and good or acceptable runs. In the MrBayes approach, we assumed the mixed+I+ Γ (5) model for all alignment sets to sample appropriate models across the substitution model space in the Bayesian MCMC analysis itself, avoiding the need for a priori model testing. In this analysis, two independent runs starting from random trees were applied, each using eight Markov chains with 40,000,000 generations (for the Rpl28 and Rpl33 sets)

or four Markov chains with 20,000,000 generations (for the other sets). Trees were sampled every 100 generations; to calculate a posterior consensus, we selected trees from the last 5,868,000-26,644,000 generations that reached stationary phase and convergence (the standard deviation of split frequencies stabilized and was less than 0.01).

In the TreeFinder approach, we applied appropriate substitution models that were chosen according to the Propose Model module in this program assuming optimized frequencies of amino acids, whereas the models used in (more)PhyML were selected according to ProtTest 3.2 (ref. 12) assuming optimization of models, branches, and topology of the tree (Table 1S). Search depth was set to 2 in TreeFinder, and the best heuristic search algorithms, NNI and SPR, in (more)PhyML were applied. Edge support was assessed by the bootstrap analysis with 1,000 replicates in each of these two programs. Additionally, we applied the Local Rearrangements-Expected Likelihood Weights method in TreeFinder and the approximate likelihood ratio test (aLRT) based on a Shimodaira-Hasegawa-like procedure in morePhyML.¹³ In PhyML-Structure, we used the EX_EHO+ Γ (5) substitution model for all alignment sets, whereas edge support was calculated by aLRT based on the χ^2 test and a Shimodaira-Hasegawa-like procedure. The minimum of these two aLRT support values is shown at selected nodes in the presented trees (Fig. 1S).

Topology tests with 10,000,000 replicates were performed in Consel v0.20 (ref. 14) to compare trees obtained in PhyloBayes under LG+ Γ (5) with alternative topologies that assumed different positions of minicircle *C. horridum* and *P. lunula* sequences (Fig. 1S). Site-wise log-likelihoods for the analyzed trees were calculated in PhyML under the best fitted substitution models found in ProtTest.

Table 1S. Applied substitution models in the analyzed alignment sets.

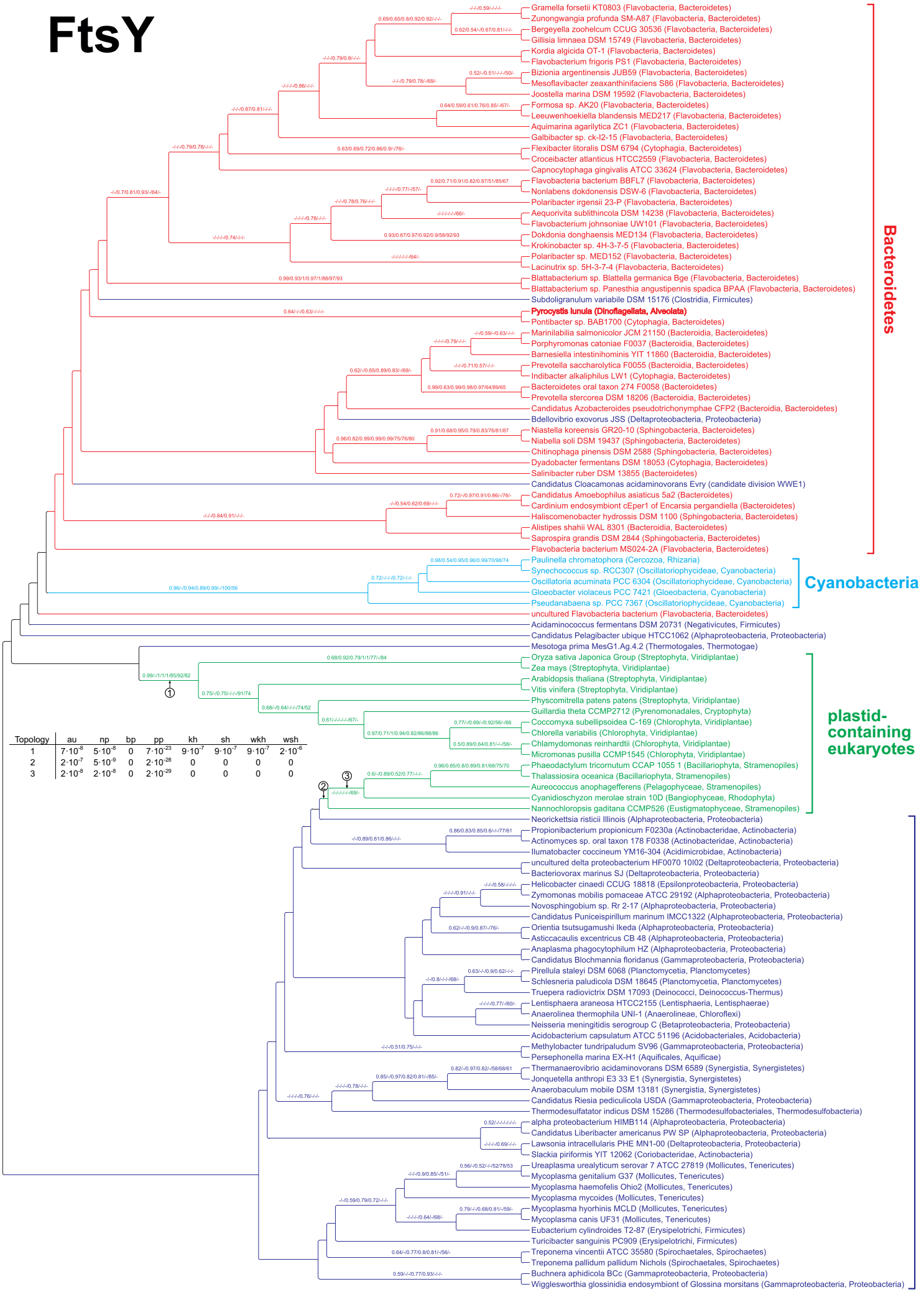
Alignment set	PhyloBayes	MrBayes	TreeFinder	(more)PhyML	PhyML-Structure
FtsY	LG+ Γ (5), CAT Poisson+ Γ (5)	mixed+I+ Γ (5)	LG+F+I+ Γ (5)	LG+ Γ (5)	EX_EHO+ Γ (5)
Rpl28	LG+ Γ (5), CAT Poisson+ Γ (5)	mixed+I+ Γ (5)	witHIV+I+ Γ (5)	LG+F+ Γ (5)	EX_EHO+ Γ (5)
Rpl33	LG+ Γ (5), CAT Poisson+ Γ (5)	mixed+I+ Γ (5)	MIX+F+ Γ (5)	LG+ Γ (5)	EX_EHO+ Γ (5)
Ycf16	LG+ Γ (5), CAT Poisson+ Γ (5)	mixed+I+ Γ (5)	LG+F+I+ Γ (5)	LG+F+ Γ (5)	EX_EHO+ Γ (5)
Ycf24	LG+ Γ (5), CAT Poisson+ Γ (5)	mixed+I+ Γ (5)	LG+F+ Γ (5)	LG+F+ Γ (5)	EX_EHO+ Γ (5)

References

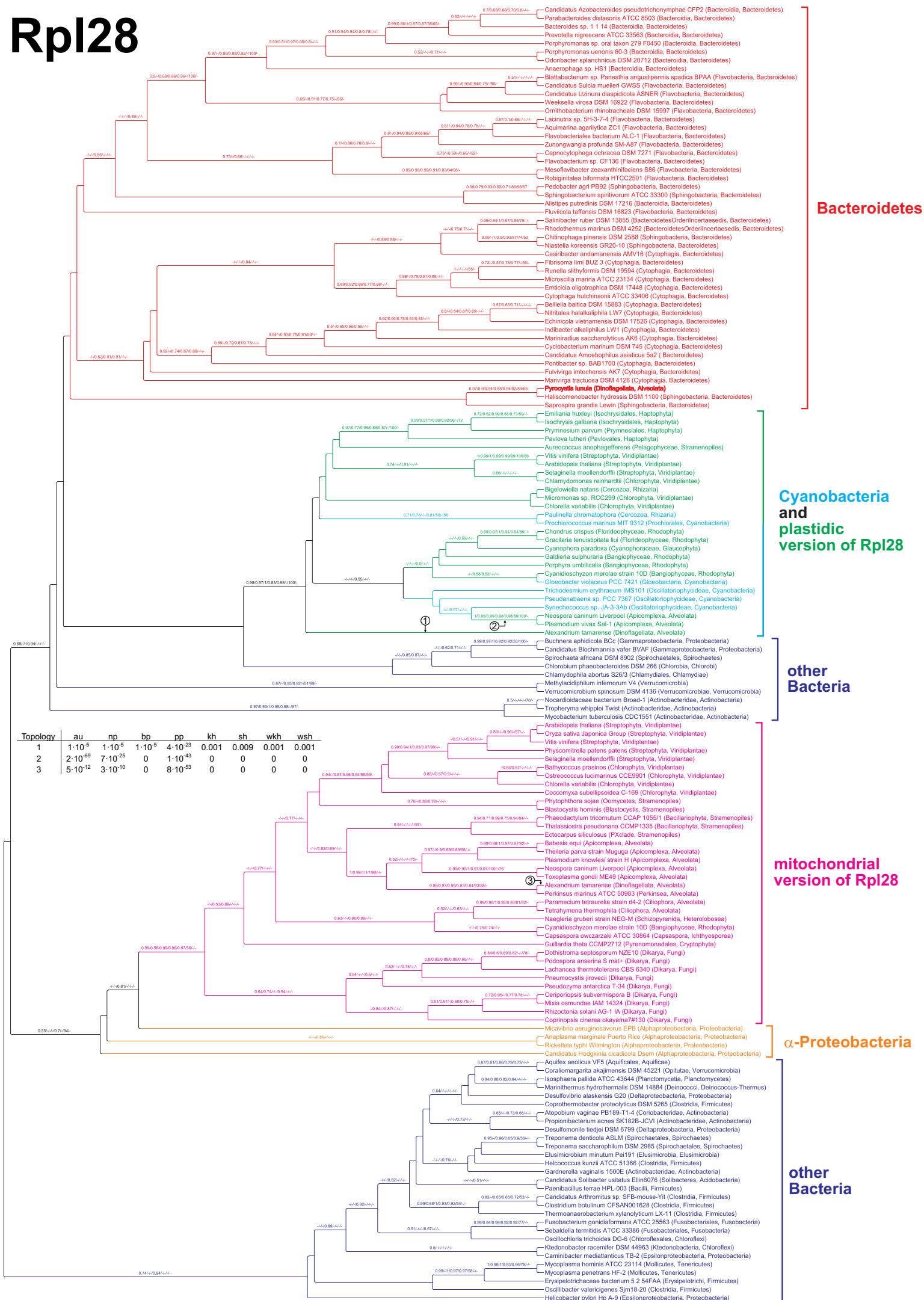
1. Chan CX, Soares MB, Bonaldo MF, Wisecaver JH, Hackett JD, Anderson DM, et al. Analysis of *Alexandrium tamarense* (Dinophyceae) genes reveals the complex evolutionary history of a microbial eukaryote. *J Phycol* 2012; 48:1130-42.
2. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, et al. CDD: a conserved domain database for protein classification. *Nucleic Acids Res* 2005; 33:D192-6.
3. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008; 9:286-98.
4. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. *Bioinformatics* 2009; 25:1189-91.
5. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000; 302:205-17.
6. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 2004; 21:1095-109.
7. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012; 61:539-42.
8. Jobb G, von Haeseler A, Strimmer K. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* 2004; 4:18.
9. Le SQ, Gascuel O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is highly beneficial. *Syst Biol* 2010; 59:277-87.
10. Criscuolo A. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. *Mol Phylogenet Evol* 2011; 61:944-8.
11. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010; 59:307-21.
12. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 2011; 27:1164-5.
13. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* 2006; 55:539-52.
14. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 2001; 17:1246-7.

Figure 1S. Phylogenetic trees for FtsY, Rpl28, Rpl33, Ycf16, and Ycf24 sequences inferred in PhyloBayes under the LG+ $\Gamma(5)$ model. Sequences localized to *Pyrocystis lunula* AF490367 and *Ceratium horridum* AF490364 minicircles appear in bold font. Numbers at nodes (in order) correspond to: posterior probabilities estimated in Phylobayes for the LG+ $\Gamma(5)$ and CAT Poisson+ $\Gamma(5)$ models as well as in MrBayes, the minimum of support values calculated by aLRT based on the χ^2 test and a Shimodaira-Hasegawa-like procedure in PhyML-Structure, support values obtained by a Shimodaira-Hasegawa-like procedure in morePhyML, PhyML bootstrap values, Local Rearrangements-Expected Likelihood Weights-support values calculated in TreeFinder, and TreeFinder bootstrap values. Values of the posterior probabilities and bootstrap percentages lower than or equal to 0.50 and 50%, respectively, were omitted or indicated by a dash “-“. Tables show the results of topology tests comparing the best topology with alternatives that assume different positions of the minicircle-encoded proteins. Topology test results are: the p-value for the approximately unbiased test (au) calculated from the multiscale bootstrap, the non-parametric bootstrap probability calculated from the multiscale bootstrap (np), the bootstrap probability calculated in the non-multiscale manner (bp), the Bayesian posterior probability calculated by the BIC approximation (pp), and the p-values of the Kishino-Hasegawa test (kh), the Shimodaira-Hasegawa test (sh), the weighted Kishino-Hasegawa test (skh), and the weighted Shimodaira-Hasegawa test (wsh).

FtsY



Rpl28



Rpl33



Bacteroidetes

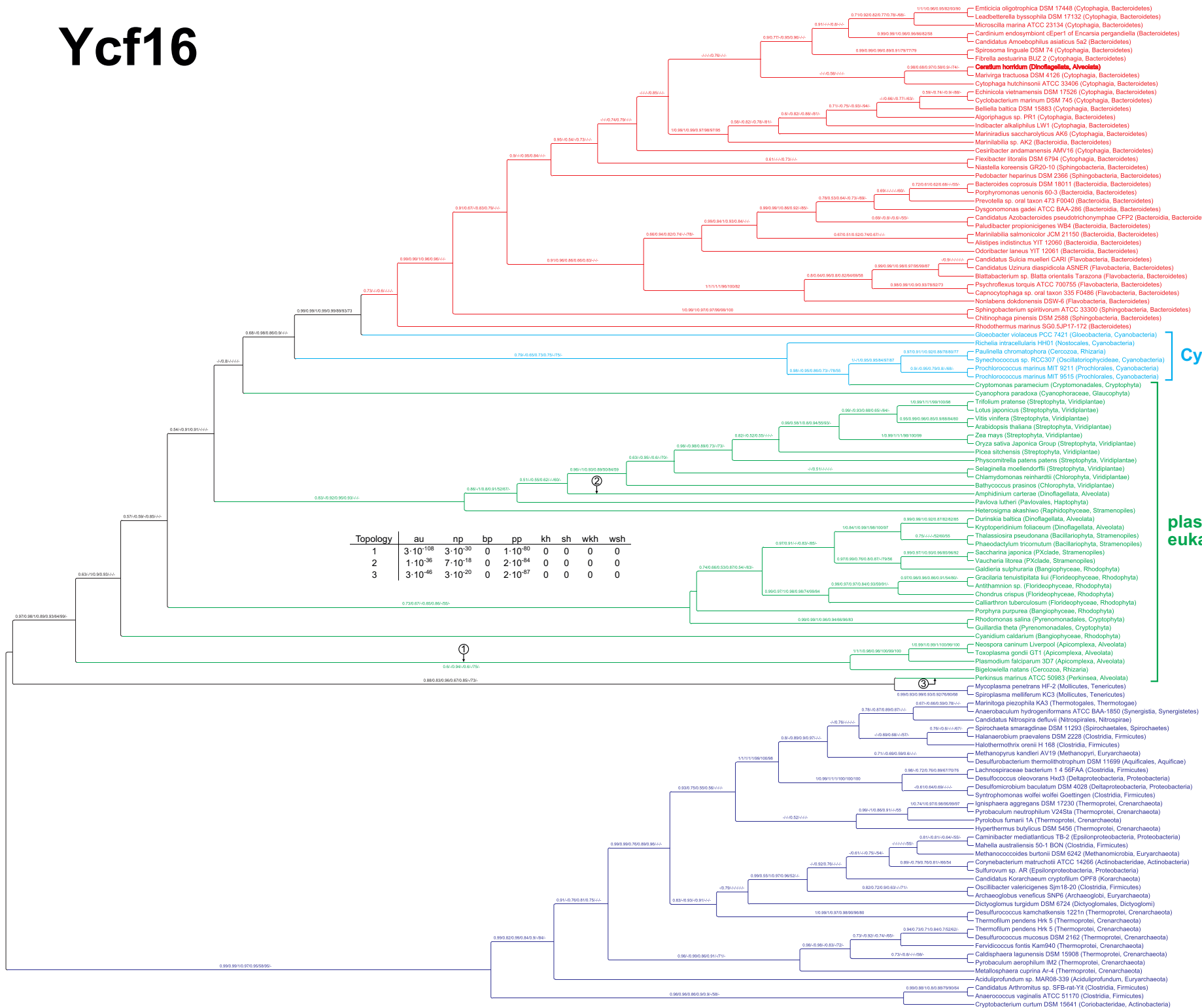
other Bacteria

Cyanobacteria and
plastid version of Rpl33

alpha-Proteobacteria and
mitochondrial version of Rpl33

other Bacteria

Ycf16



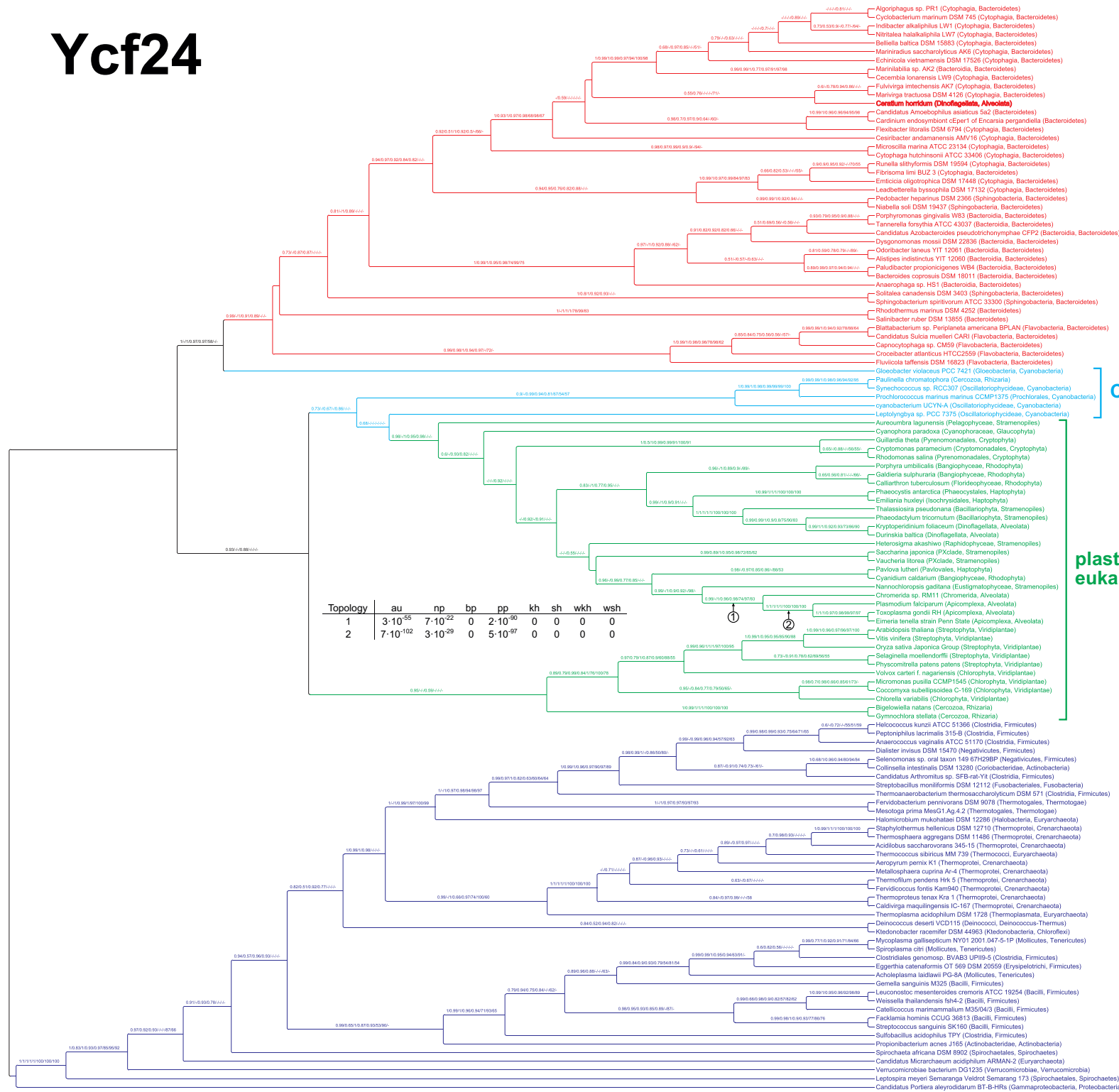
Bacteroidetes

Cyanobacteria

plastid-containing eukaryotes

other Bacteria and Archaea

Ycf24



Bacteroidetes

Cyanobacteria

plastid-containing eukaryotes

other Bacteria and Archaea