Supplementary Information for:
Phylogenomic Insights into the Origin of Primary Plastids

Iker Irisarri, Jürgen F. H. Strassert, Fabien Burki

Supplementary Table 1. UFBoot support for monophyly of Archaeplastida (Arch.) from all four source datasets, along with the support for Glaucophyta (Glau.), Rhodophyta (Rhod.), Chloroplastida (Chlo.), and SAR. The support from subsets of common genes and taxa are also shown, where datasets should be interpreted as follows: the gene-overlap dataset BAU-ALL refers for the subset of genes common to all four datasets using the BAU alignments; BAU-BRO refers to the subset of shared genes between BAU and BRO using the BAU alignments and vice versa for BRO-BAU. All analyses were performed with IQ-TREE under BIC-selected models. NA refers to cases where a clade was represented by a single taxon.

Original datasets

Dataset	Model	SAR	Arch.	Glau.	Rhod.	Chlo.	No. taxa	No. genes
BAU	LG+Γ4	100	0	100	100	100	57	108
BAU	LG+C40+F+Γ4	100	0	100	100	100	57	180
BRO	LG+F+Γ4	100	0	100	100	100	68	159
BRO	LG+C40+F+Γ4	100	0	100	100	100	68	159
BUR	LG+F+Γ4	100	0	100	100	100	150	250
BUR	LG+C60+F+Γ4	100	0	100	100	100	150	250
KAT	LG+F+Γ4	100	95	100	100	100	231	150
KAT	LG+C60+F+Γ4	100	91	100	100	100	231	150

"Gene overlap" datasets

Dataset	Model	SAR	Arch.	Glau.	Rhod.	Chlo.	No. taxa	No. genes
BAU-ALL	LG+Γ4	100	0	100	100	100	56	57
BAU-ALL	LG+C60+F+Γ4	100	0	100	100	100	56	57
BAU-BRO	LG+Γ4	100	0	100	100	100	56	84
BAU-BUR	LG+Γ4	100	0	100	100	100	56	89
BAU-KAT	LG+Γ4	100	0	100	100	100	56	66
BRO-ALL	LG+Γ4	100	0	100	100	100	68	57
BRO-ALL	LG+C60+F+Γ4	100	0	100	100	100	68	57
BRO-BAU	LG+Γ4	100	0	100	100	100	68	84
BRO-BUR	LG+F+Γ4	100	0	100	100	100	68	121
BRO-KAT	LG+Γ4	100	0	100	100	100	68	75
BUR-ALL	LG+Γ4	100	0	100	100	100	150	57
BUR-ALL	LG+C60+F+Γ4	100	0	100	100	100	150	57
BUR-BAU	LG+Γ4	100	0	100	100	100	150	89
BUR-BRO	LG+Γ4	100	0	100	100	100	150	121
BUR-KAT	LG+F+Γ4	100	0	100	100	100	150	79
KAT-ALL	LG+Γ4	73	0	100	84	100	227	57
KAT-ALL	LG+C60+F+Γ4	73	0	100	84	100	227	57
KAT-BAU	LG+Γ4	0	0	0	100	100	229	66
KAT-BRO	LG+Γ4	100	0	100	99	100	231	75
KAT-BUR	LG+Γ4	100	0	96	66	100	229	79

"Taxon-overlap" datasets

Dataset	Model	SAR	Arch.	Glau.	Rhod.	Chlo.	No. taxa	No.genes
BAU-ALL	LG+F+Γ4	100	0	NA	100	NA	20	20
BAU-ALL	LG+C60+F+Γ4	100	0	NA	100	NA	20	20
BAU-BRO	LG+Γ4	100	0	NA	100	100	28	29
BAU-BUR	LG+Γ4	100	0	NA	100	100	33	35
BAU-KAT	LG+Γ4	100	0	100	100	100	44	48
BRO-ALL	LG+F+Γ4	100	0	NA	100	NA	20	20
BRO-ALL	LG+C60+F+Γ4	100	0	NA	100	NA	20	20
BRO-BAU	LG+F+Γ4	100	0	100	100	NA	29	23
BRO-BUR	LG+F+Γ4	100	0	NA	100	100	41	41
BRO-KAT	LG+F+Γ4	100	0	100	100	100	57	57
BUR-ALL	LG+F+Γ4	100	0	NA	100	NA	20	20
BUR-ALL	LG+C60+F+Γ4	100	0	NA	100	NA	20	20
BUR-BAU	LG+F+Γ4	100	0	100	100	100	34	35
BUR-BRO	LG+F+Γ4	100	0	NA	100	100	39	41
BUR-KAT	LG+F+Γ4	100	0	NA	100	100	64	66
KAT-ALL	LG+F+Γ4	100	0	NA	100	NA	20	20
KAT-ALL	LG+C60+F+Γ4	92	0	NA	100	NA	20	20
KAT-BAU	LG+F+Γ4	100	0	100	100	100	48	48
KAT-BRO	LG+F+Γ4	100	51	100	100	100	57	57
KAT-BUR	LG+F+Γ4	100	0	NA	100	100	66	66

Supplementary Table 2. Number of genes supporting tree topologies where Archaeplastida were monophyletic or non-monophyletic.

Dataset	Non-monoph. Archaeplastida	Monophyly Archaeplastida	Total	P (χ^2-test)
BAU	60	48	108	0.2482
BRO	93	66	159	**0.0323**
BUR	143	107	250	**0.0228**
KAT	76	75	151	0.9351

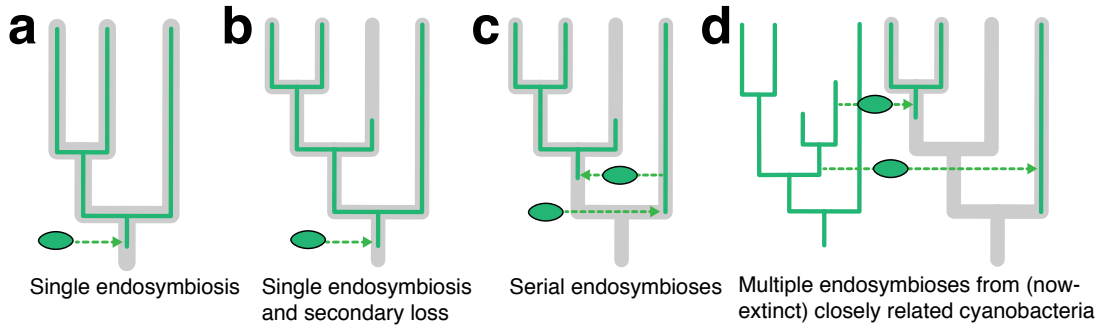
Supplementary Table 3. Alignment completeness metrics (*sensu* Wong et al. 2020) for the four original datasets and the new combined datasets, obtained with AliStat. Abbreviations: number of taxa (Ntax), positions (Npos), sequence pairs (SeqPairs), completeness of alignment matrix (Ca), row or taxa completeness (Cr), column or alignment position completeness (Cc) and proportion of completely specified homologous sites (Cij).

Original datasets

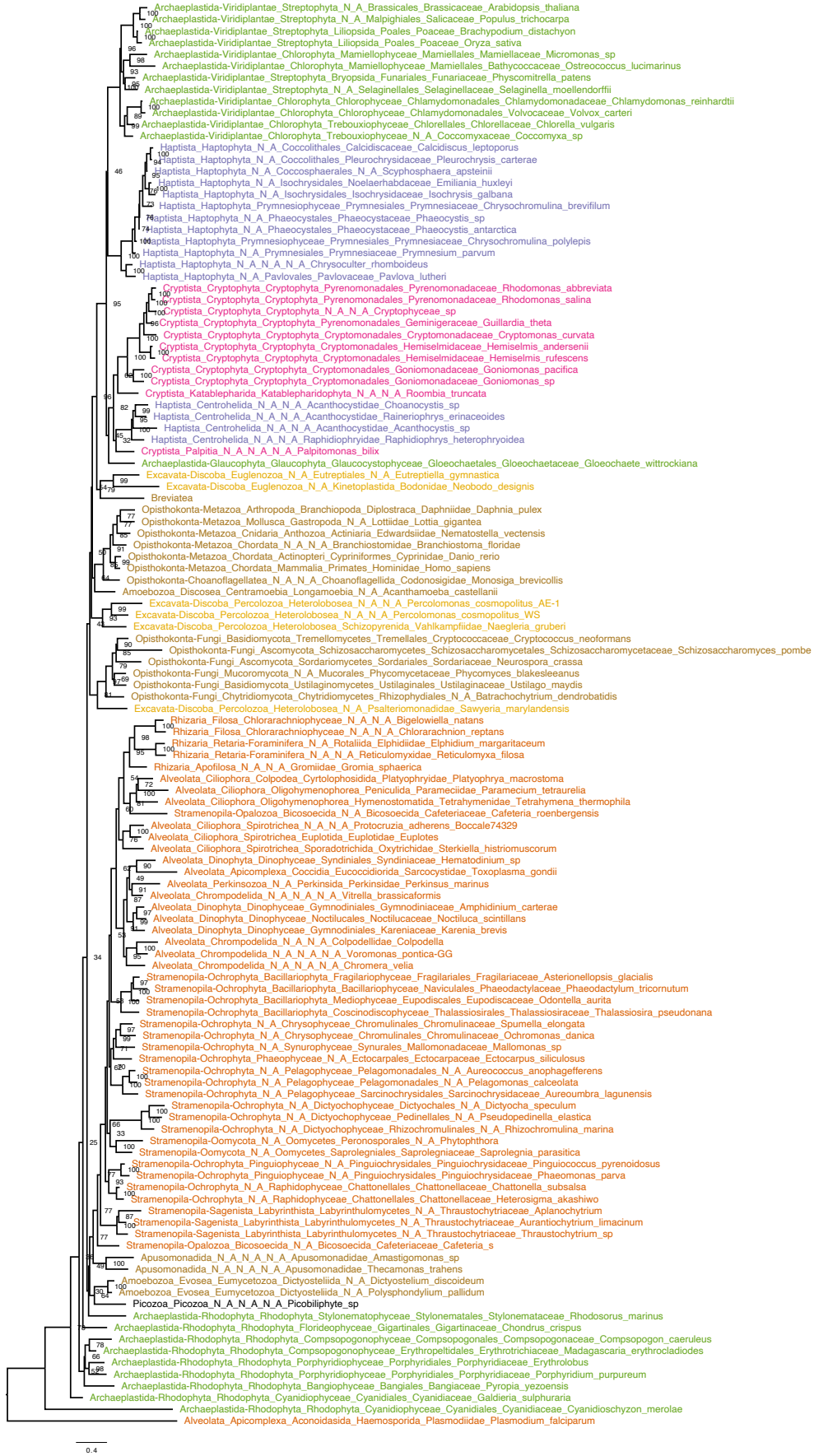
Dataset	Ntax	Npos	SeqPairs	Ca	Cr_max	Cr_min	Cc_max	Cc_min	Cij_max	Cij_min
BAU	57	15,392	1,596	0.84	1.00	0.23	1.00	0.70	1.00	0.09
BRO	68	43,615	2,278	0.61	0.99	0.19	0.97	0.01	0.98	0.06
BUR	150	55,554	11,175	0.79	0.98	0.14	0.96	0.43	0.96	0.03
KAT-PROTS	231	34,991	26,565	0.58	1.00	0.01	0.75	0.50	1.00	0.00
KAT-18S	231	1,355	26,565	0.89	1.00	0.00	0.93	0.74	1.00	0.00

Combined datasets

Dataset	Ntax	Npos	SeqPairs	Ca	Cr_max	Cr_min	Cc_max	Cc_min	Cij_max	Cij_min
COMB-BMGE	346	75,152	59,685	0.64	0.98	0.01	0.84	0.31	0.95	0.00
COMB-BMGE-101	101	75,152	5,050	0.76	0.98	0.13	0.97	0.30	0.95	0.02
COMB-DIVPART	346	160,667	59,685	0.37	0.65	0.005	0.84	0.01	0.62	0.00
COMB-DIVPART-101	101	150,904	5,050	0.47	0.69	0.06	0.97	0.01	0.65	0.01
COMB-UNTRIM	346	202,066	59,685	0.34	0.59	0.004	0.84	0.01	0.56	0.00
COMB-UNTRIM-101	101	196,609	5,050	0.42	0.61	0.06	0.97	0.01	0.56	0.01

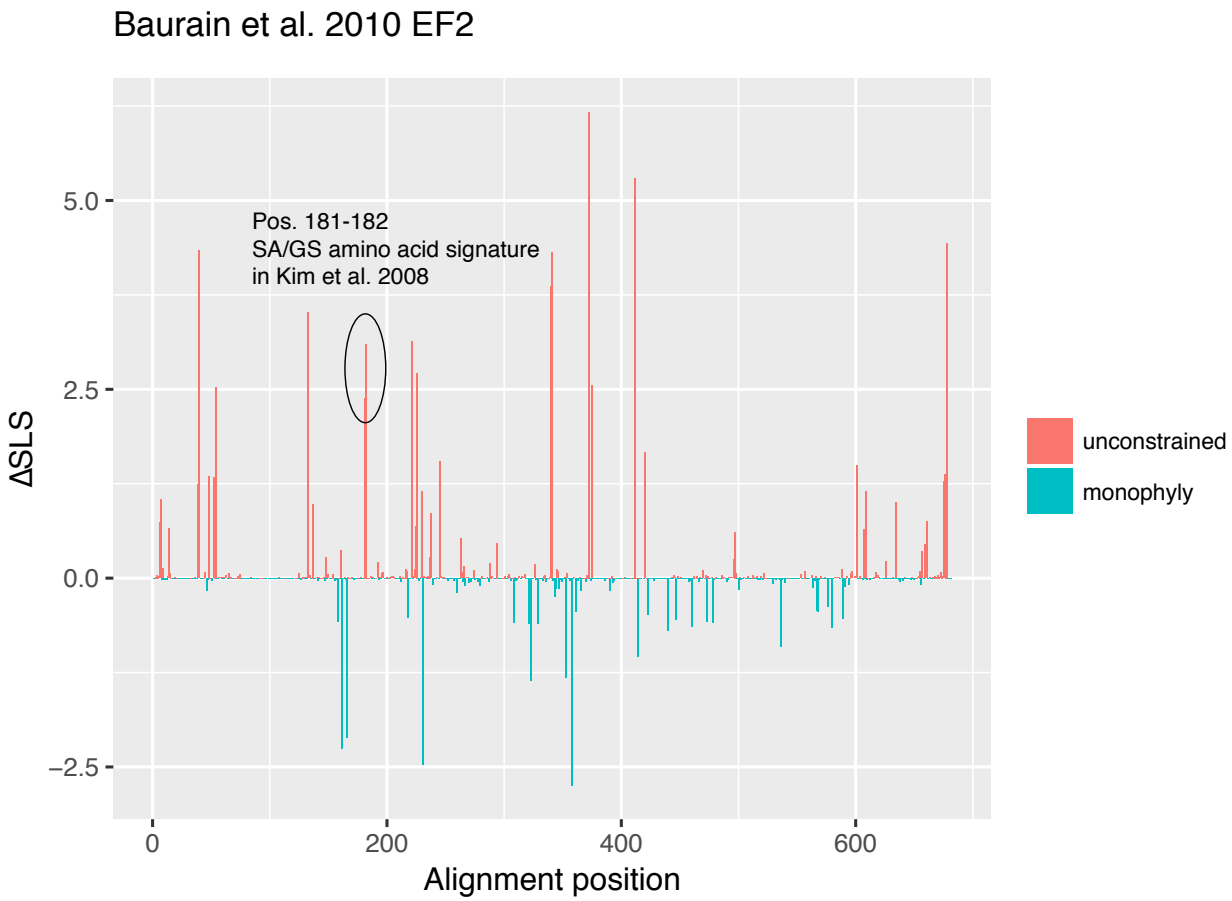


Supplementary Fig 1. Possible biological scenarios for the origin of plastids (modified from Mackiewicz and Gagat, 2014).



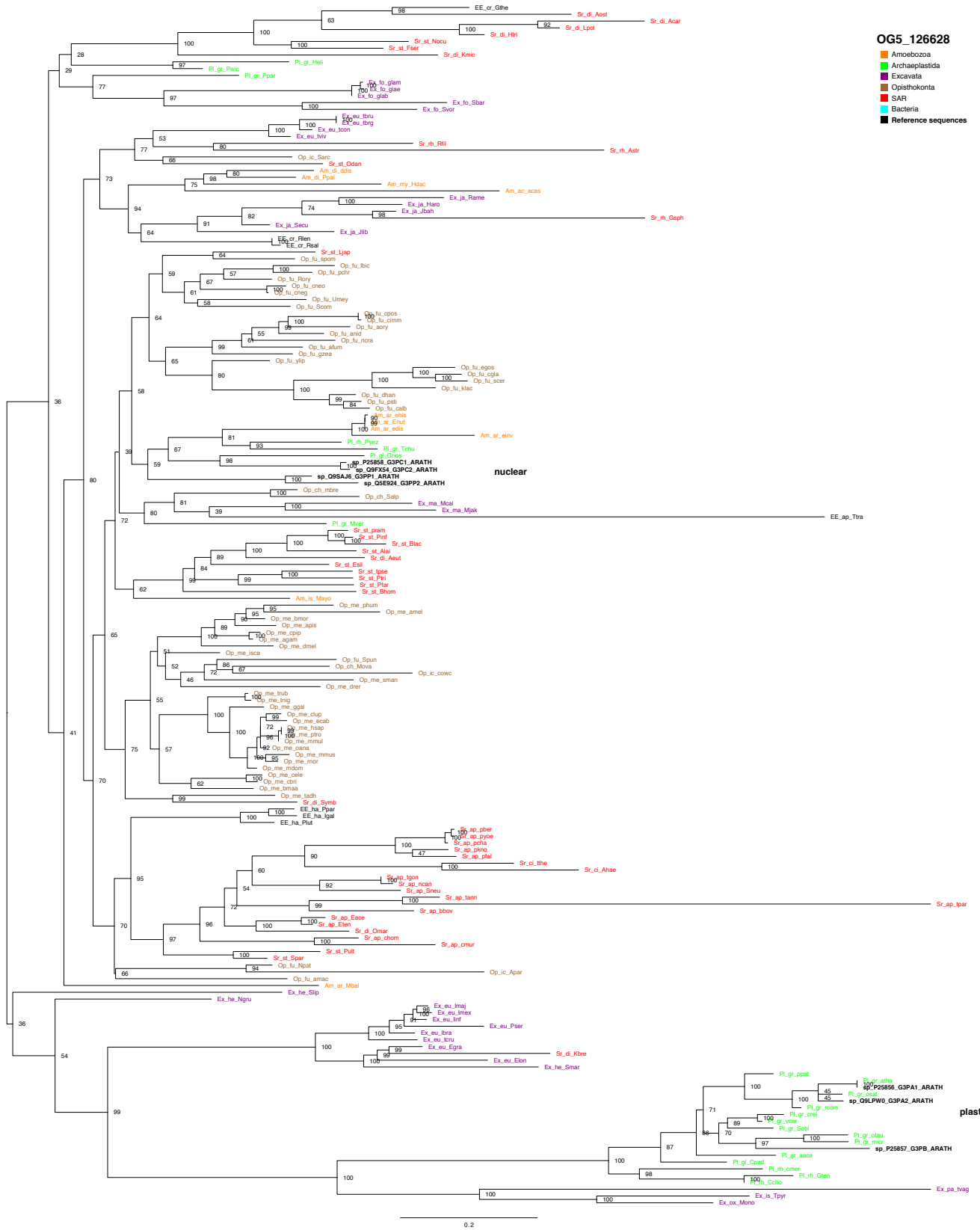
Supplementary Fig. 2. Gene tree of the *UBA3* gene from the BUR dataset (IQ-TREE). Numbers at nodes are UFBoot supports.

Supplementary Fig. 3. Gene tree of the *EF2* gene from the BAU dataset (IQ-TREE). Numbers at nodes are UFBoot supports.

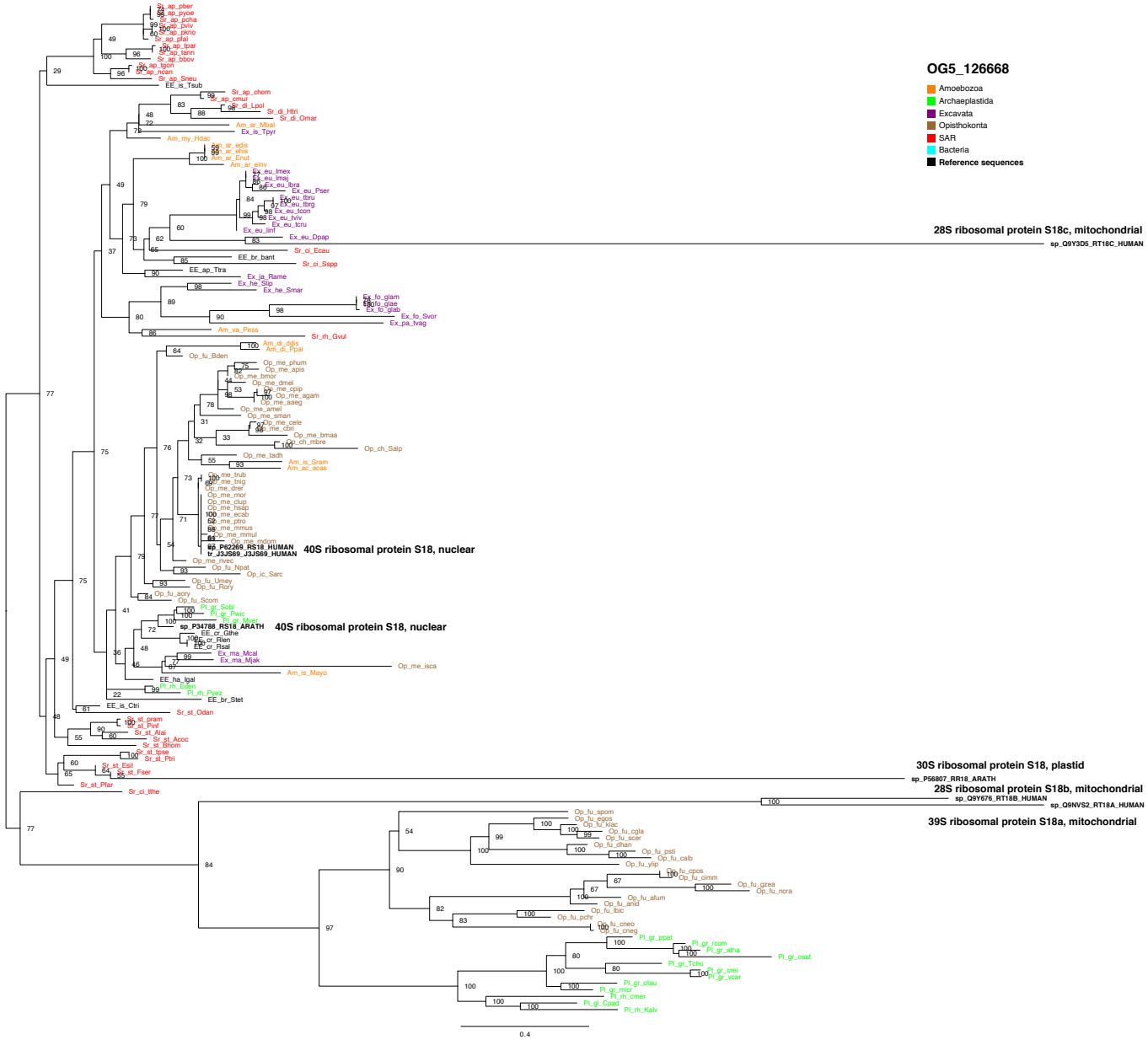


Supplementary Fig. 4. Site-wise likelihood scores (ΔSLS) for *EF2* in BAU. The SA/GS signature identified by Kim and Graham (2008) are highlighted (alignment positions 181-182).

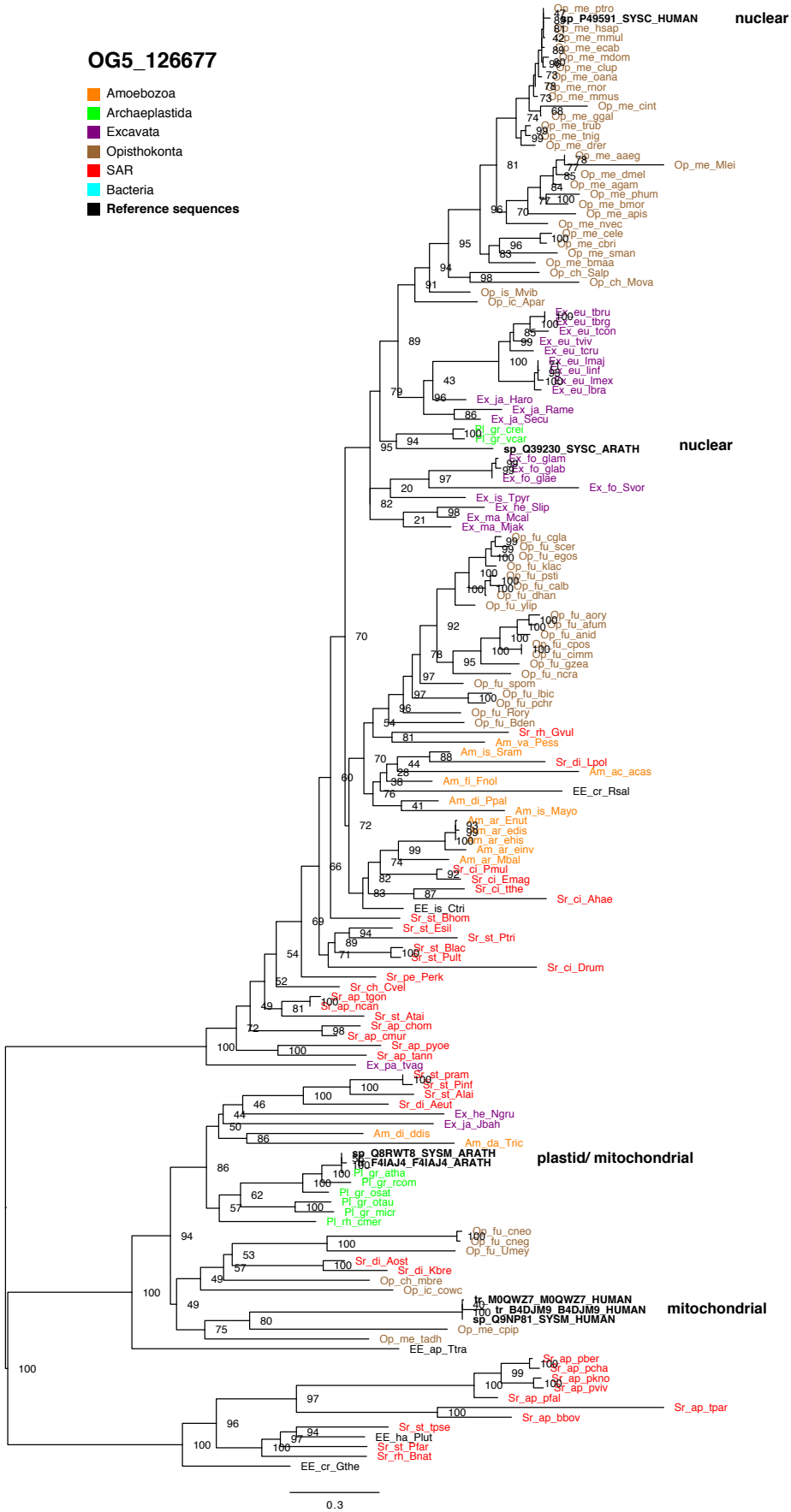
Supplementary Fig. 5. Reanalysis of the BAU dataset after removal of the *EF2* gene (IQ-TREE, LG+C40+F+Γ). Numbers at nodes are UFBoot / SH-like aLRT supports.



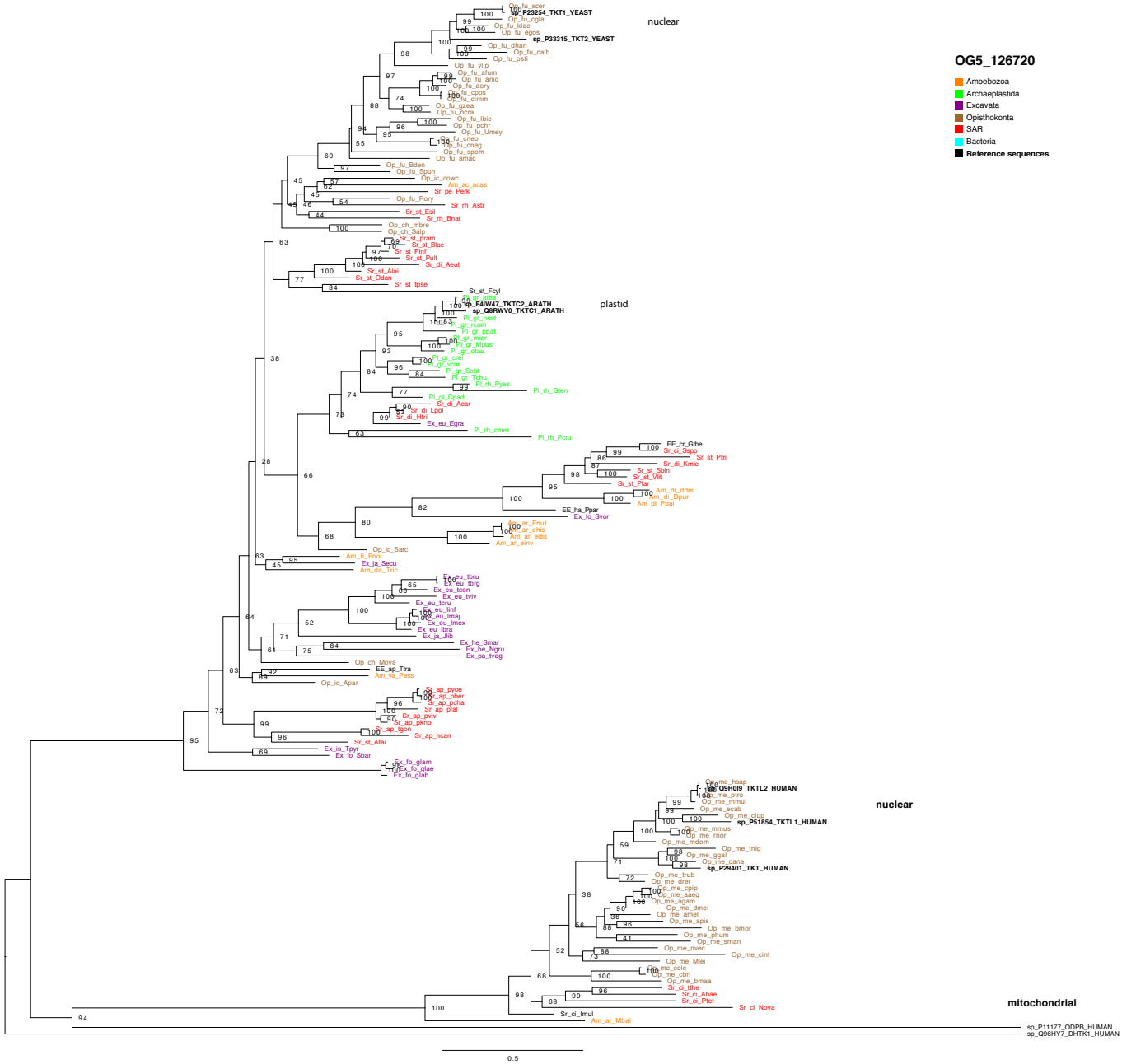
Supplementary Fig. 6. Gene tree of the *G3PC* gene (OG5_126628) from the KAT dataset (IQ-TREE). Numbers at nodes are UFBoot supports.



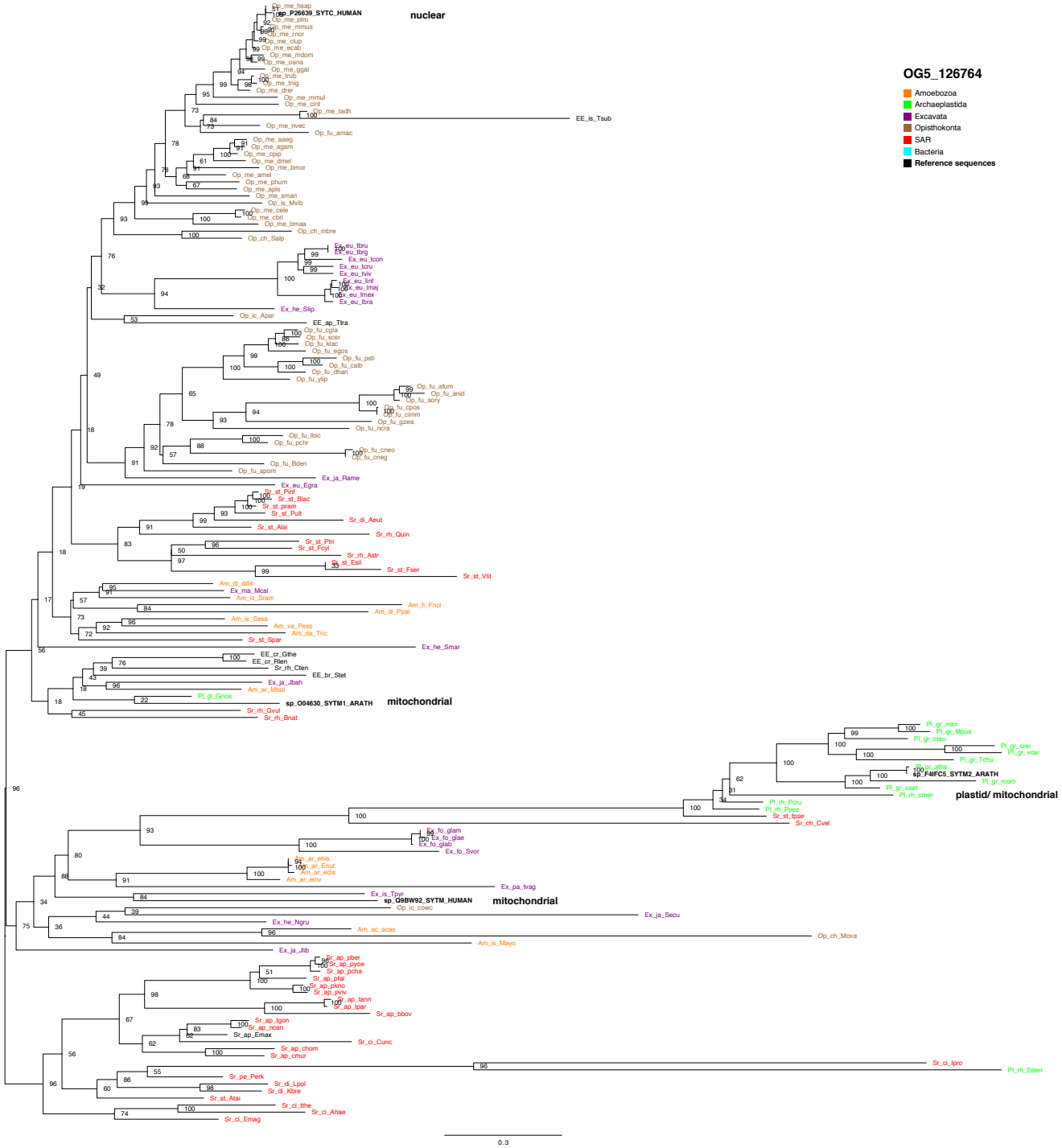
Supplementary Fig. 7. Gene tree of the *RPS18* gene (OG5_126668) from the KAT dataset (IQ-TREE). Numbers at nodes are UFBoot supports.



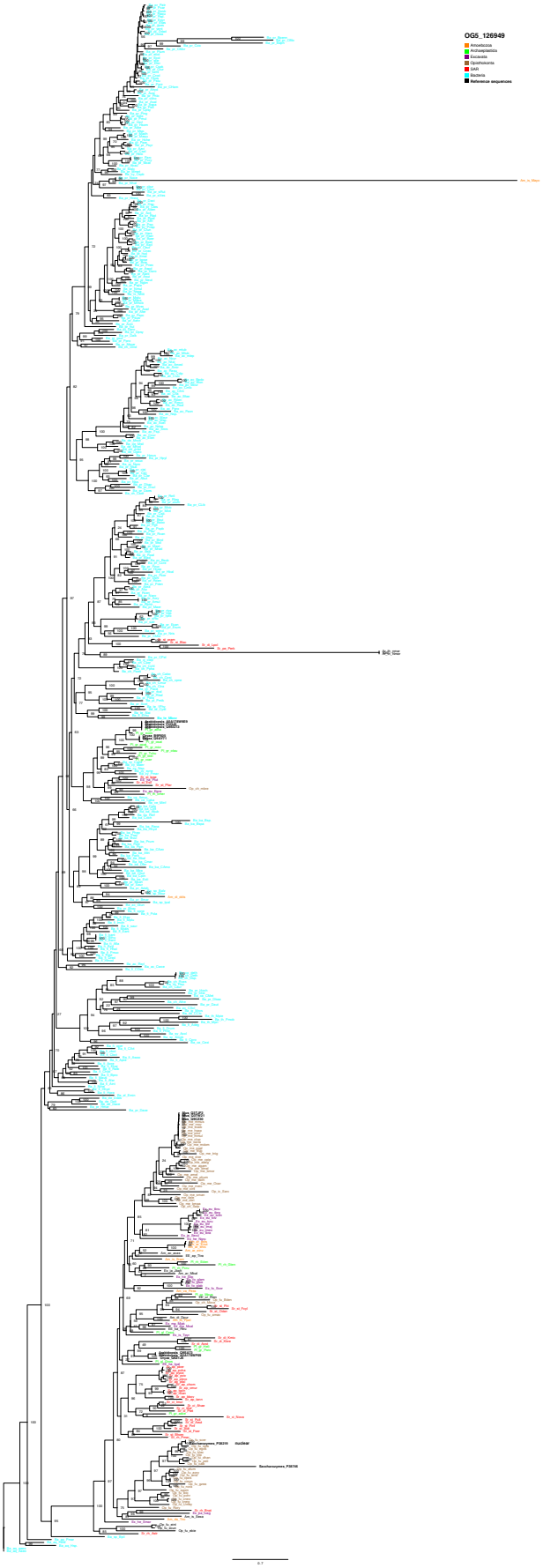
Supplementary Fig. 8. Gene tree of the *SYSC* gene (OG5_126677) from the KAT dataset (IQ-TREE). Numbers at nodes are UFBoot supports.



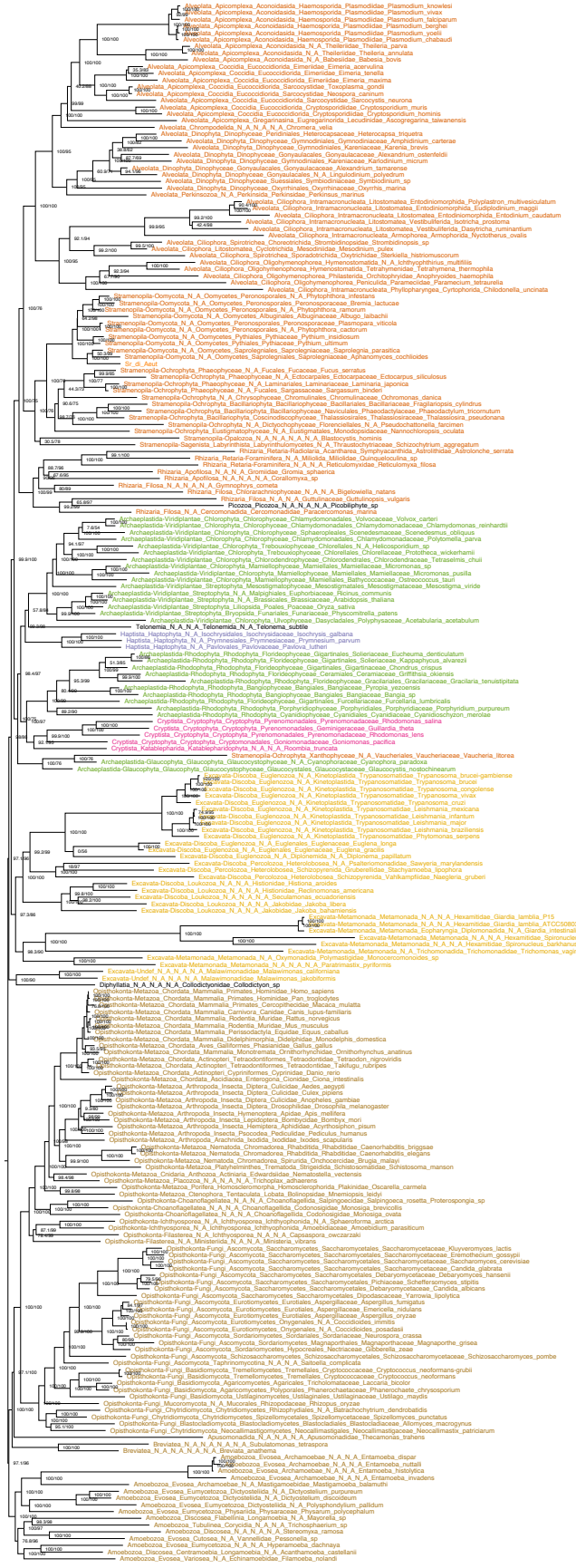
Supplementary Fig. 9. Gene tree of the *TKT1* gene (OG5_126720) from the KAT dataset (IQ-TREE). Numbers at nodes are UFBoot supports.



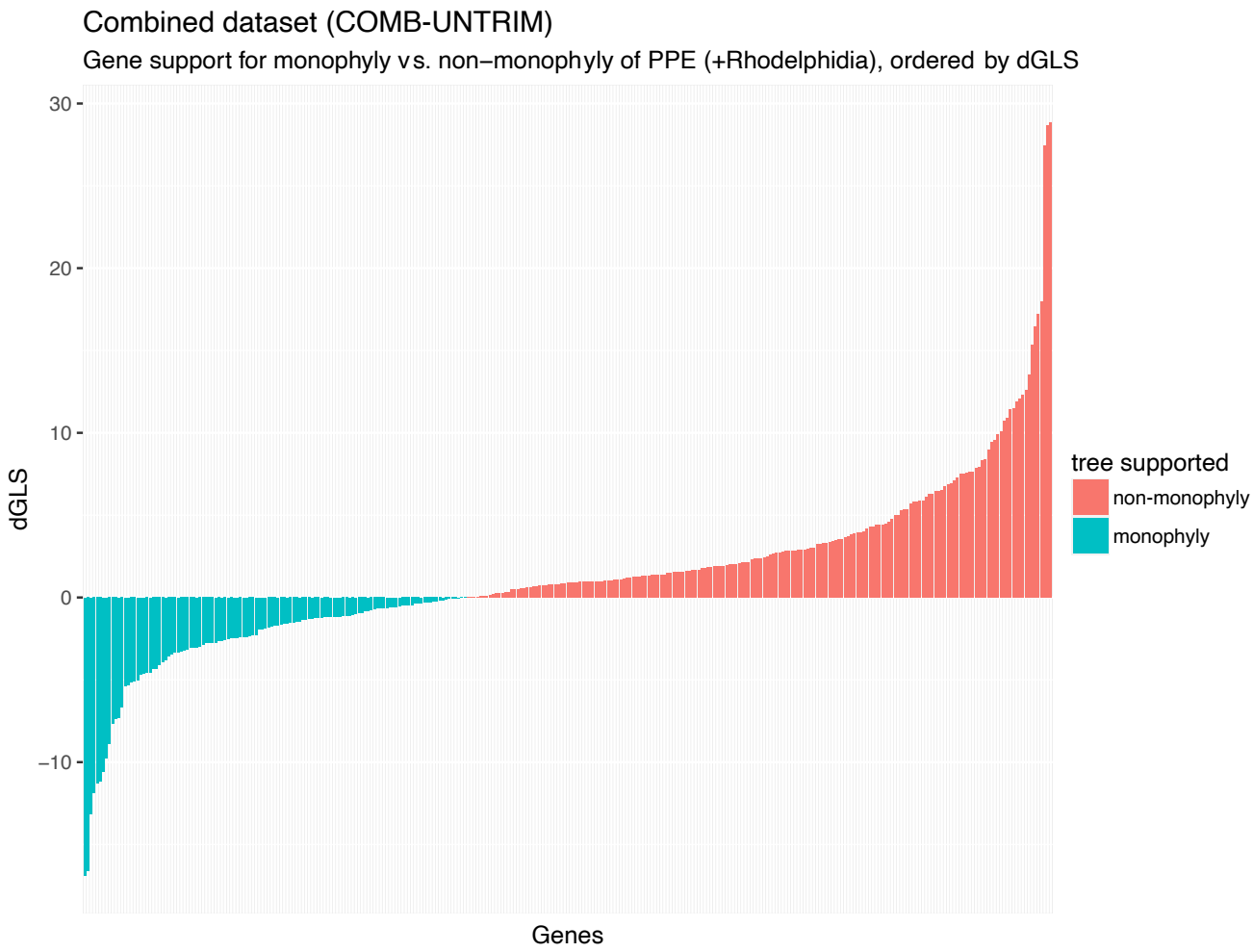
Supplementary Fig. 10. Gene tree of the *SYTC* gene (OG5_126764) from the KAT dataset (IQ-TREE). Numbers at nodes are UFBoot supports.



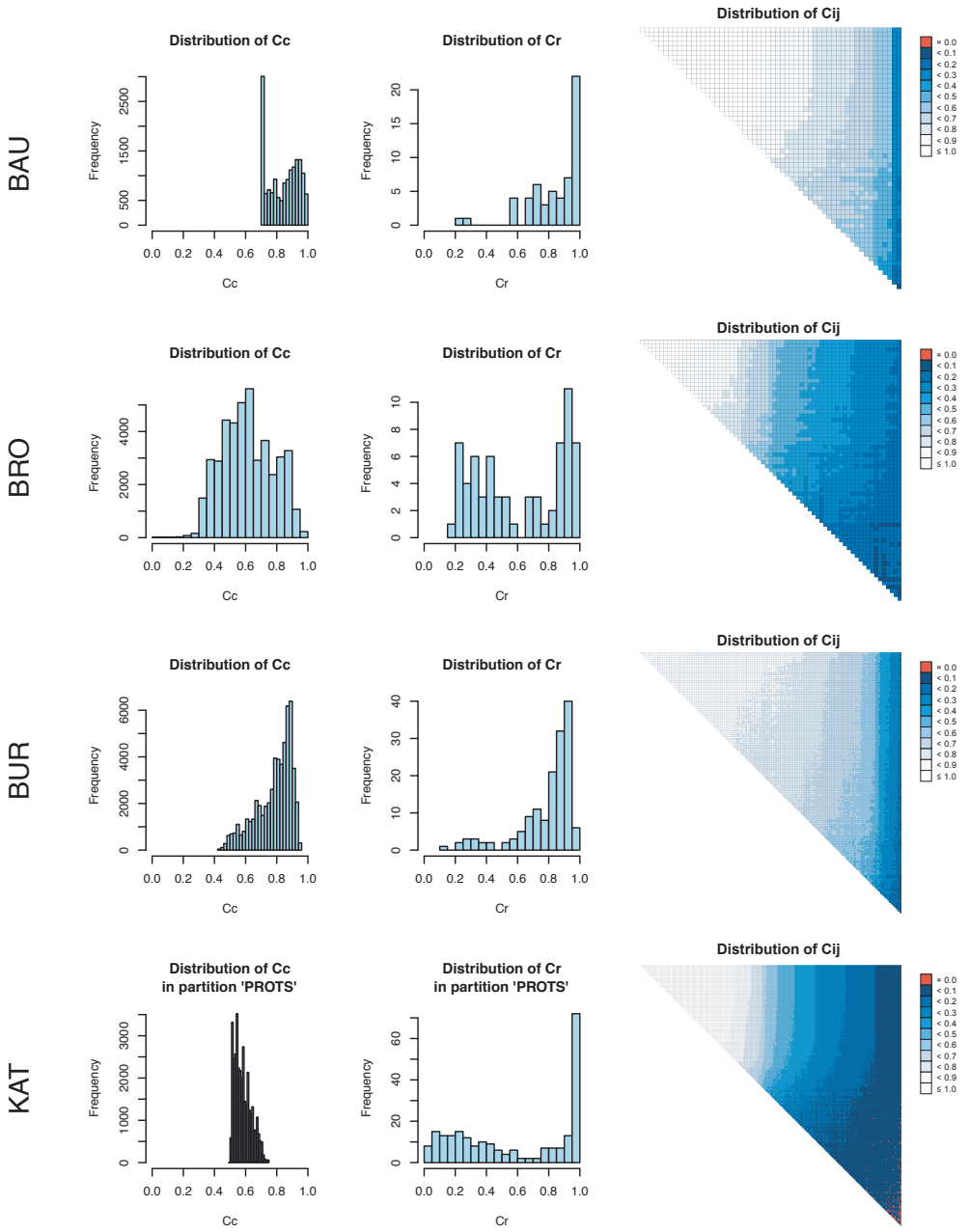
Supplementary Fig. 11. Gene tree of the *OLA1* gene (OG5_126949) from the KAT dataset (IQ-TREE). Numbers at branches are UFBoot supports.



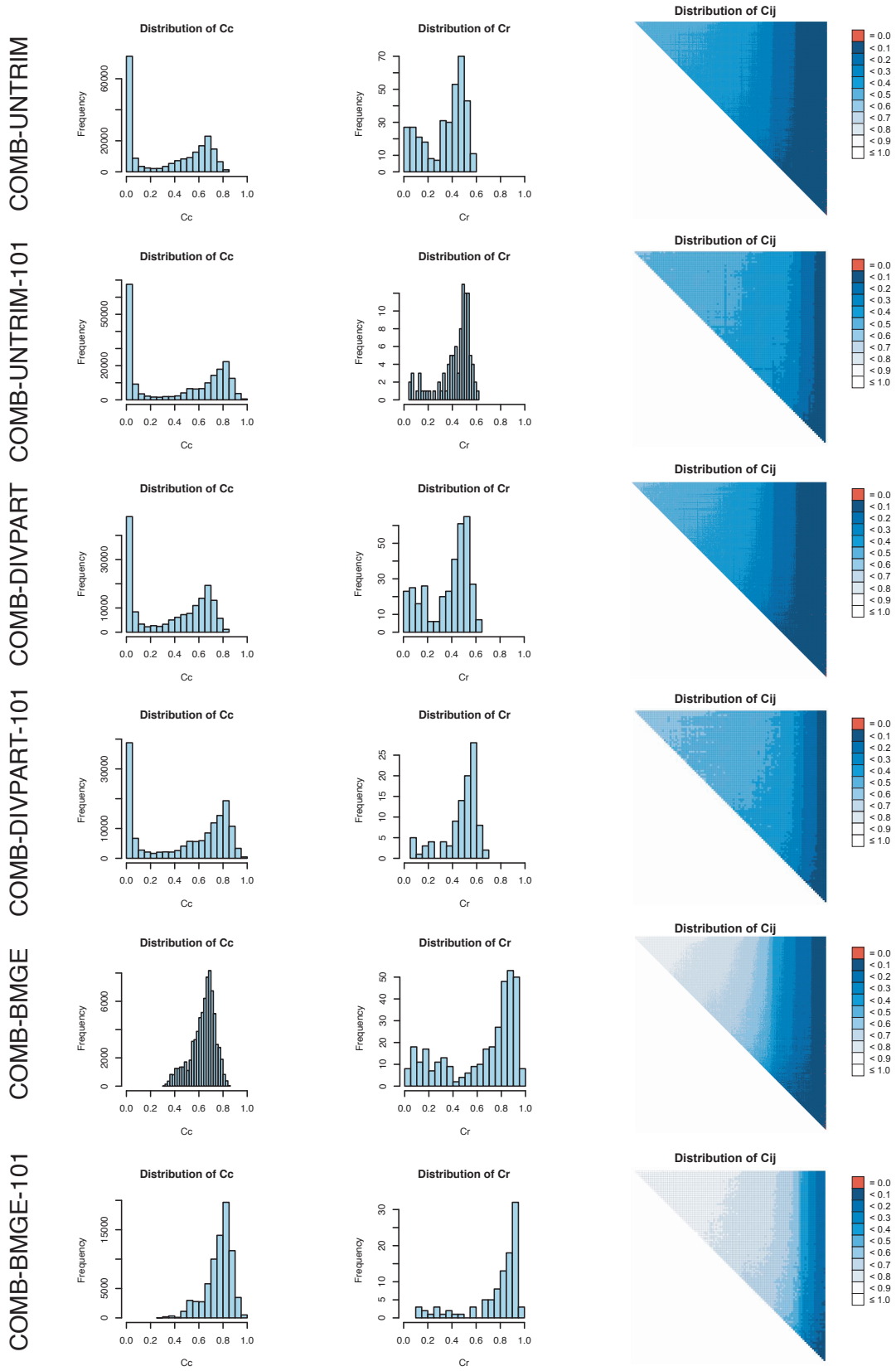
Supplementary Fig. 12. Reanalysis of the KAT dataset (IQ-TREE LG+C60+F+Γ4) after removal of paralog sequences from outlier loci. Numbers at nodes are UFBoot / SH-like aLRT supports.



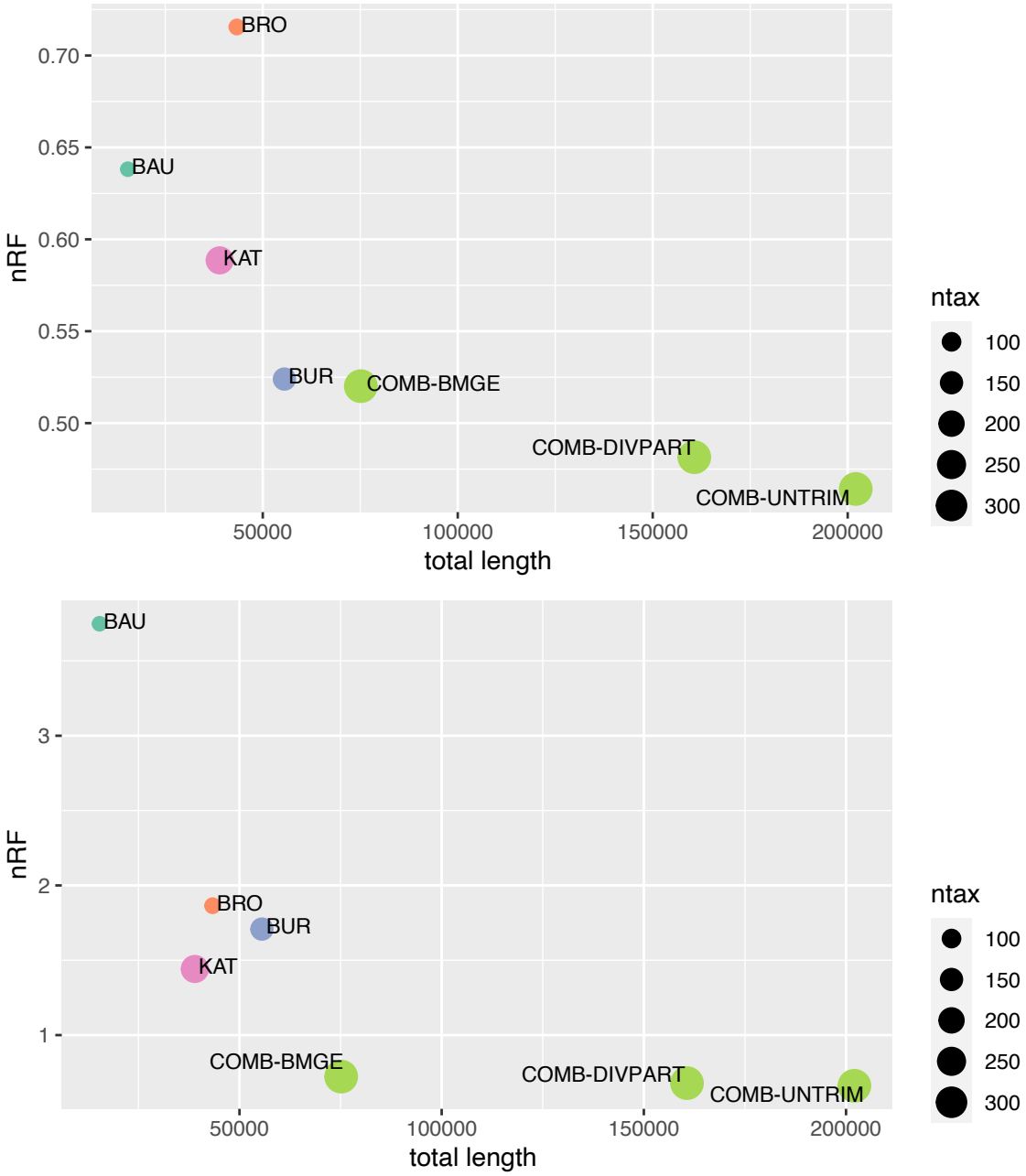
Supplementary Fig. 13. Outlier detection of the combined dataset (COMB-UNTRIM). Gene-wise log-likelihood score differences (dGLS) in support (blue) or against (red) the monophyly of primary photosynthetic eukaryotes (PPE, incl. Rhodelphidia). Calculations follow Shen et al. (2017).



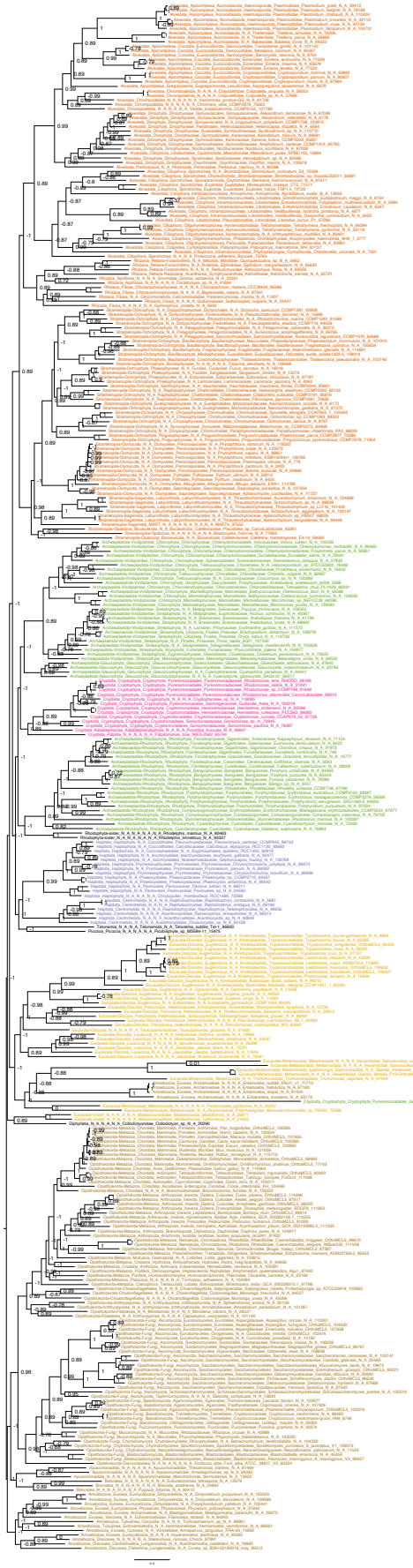
Supplementary Fig. 14. Distribution of alignment completeness metrics (*sensu* Wong et al. 2020) for the four original datasets.



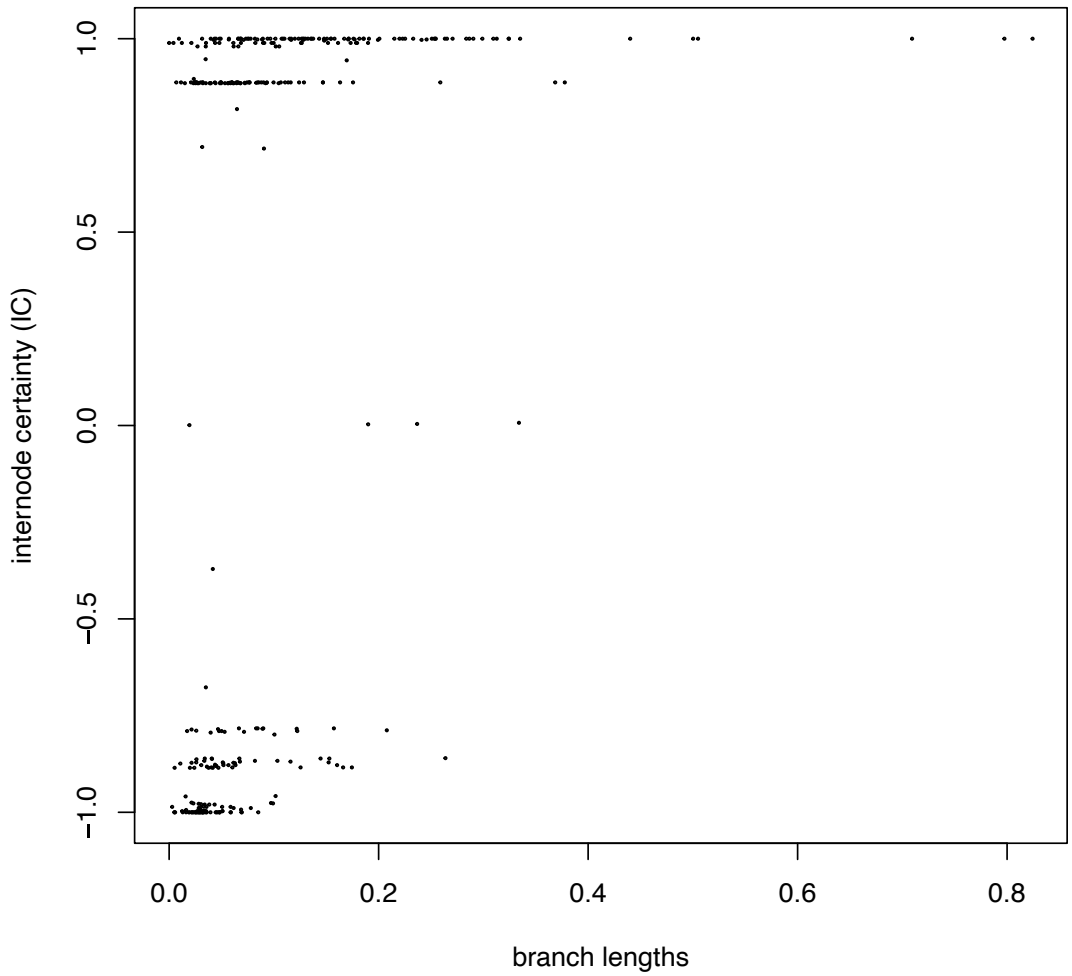
Supplementary Fig. 15. Distribution of alignment completeness metrics (*sensu* Wong et al. 2020) for the combined datasets.



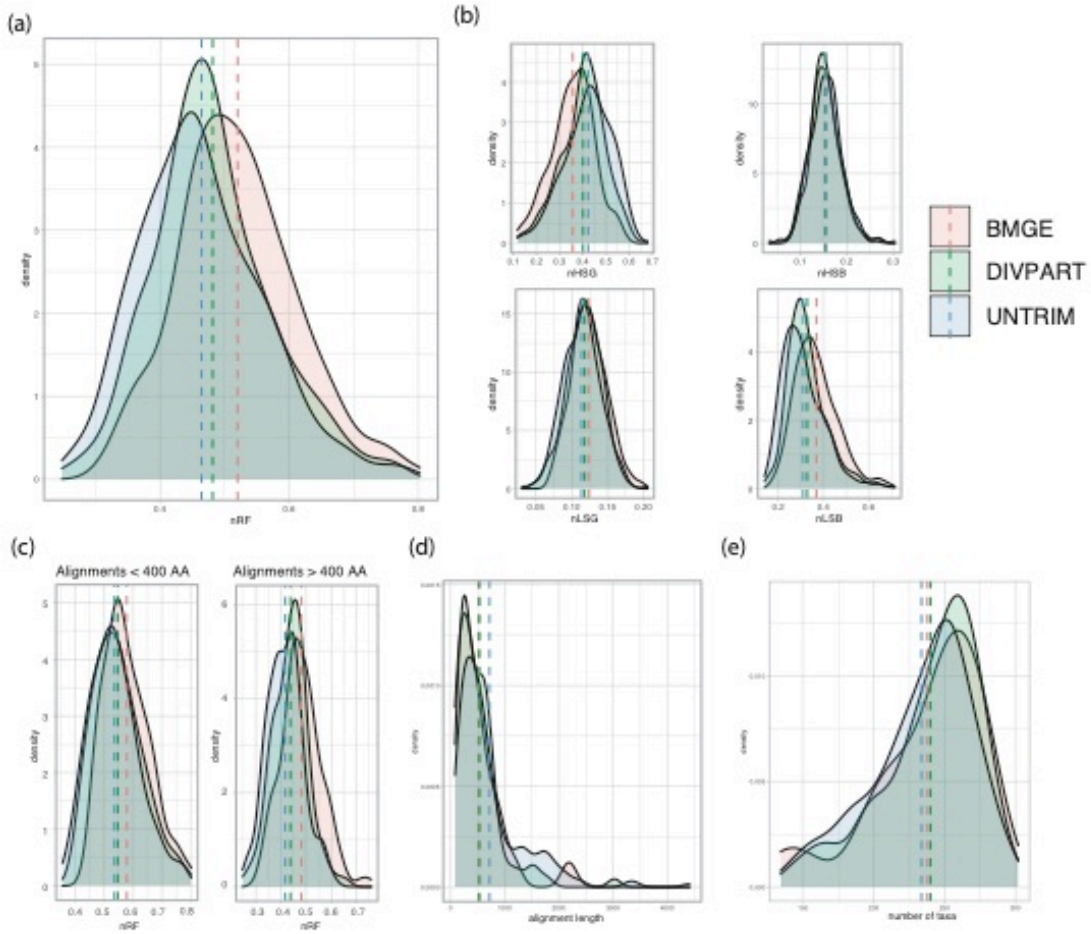
Supplementary Fig. 16. Comparison of phylogenomic datasets with respect to internal congruence among gene trees and concatenated trees (proxy for phylogenetic signal) and dataset size (total length in aligned amino acids and number of taxa). Topological distances are measured as normalized Robinson-Foulds distances (nRF). The lower graph shows nRF corrected by gene length.



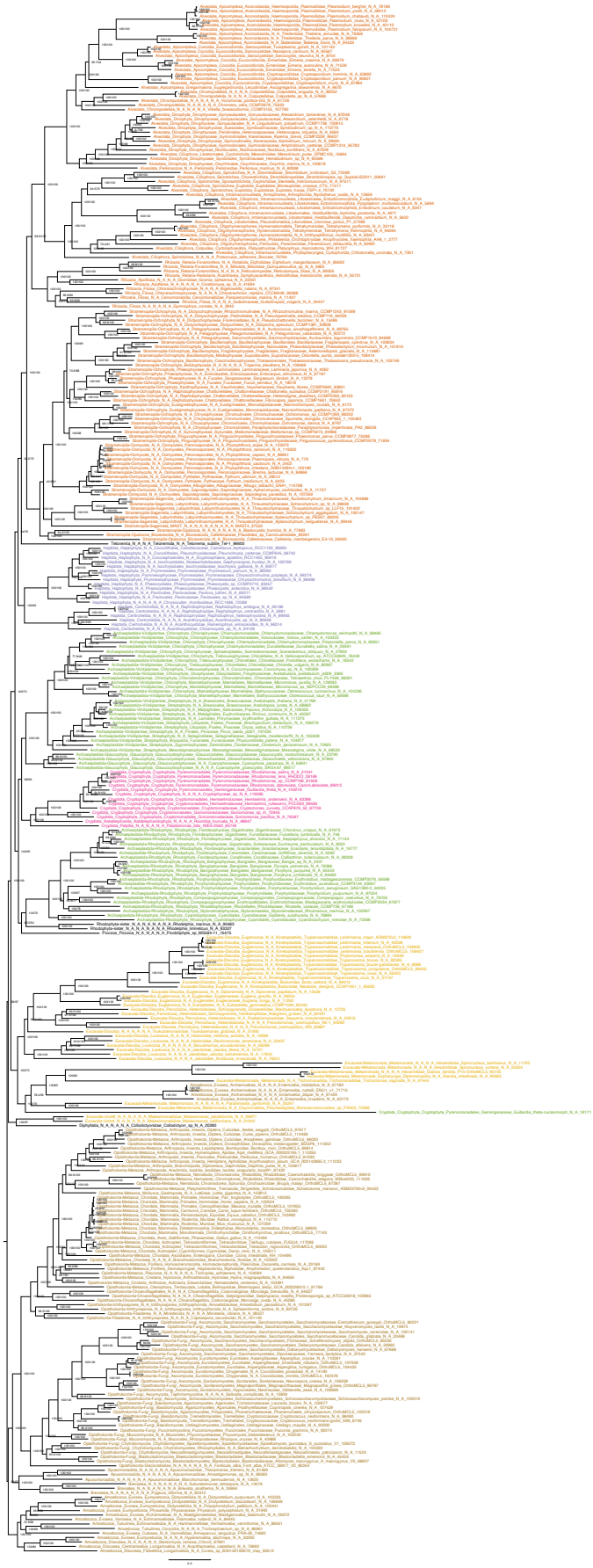
Supplementary Fig. 17. Internode certainty (IC) scores on the COMB-UNTRIM dataset (probabilistic correction for incomplete gene trees).



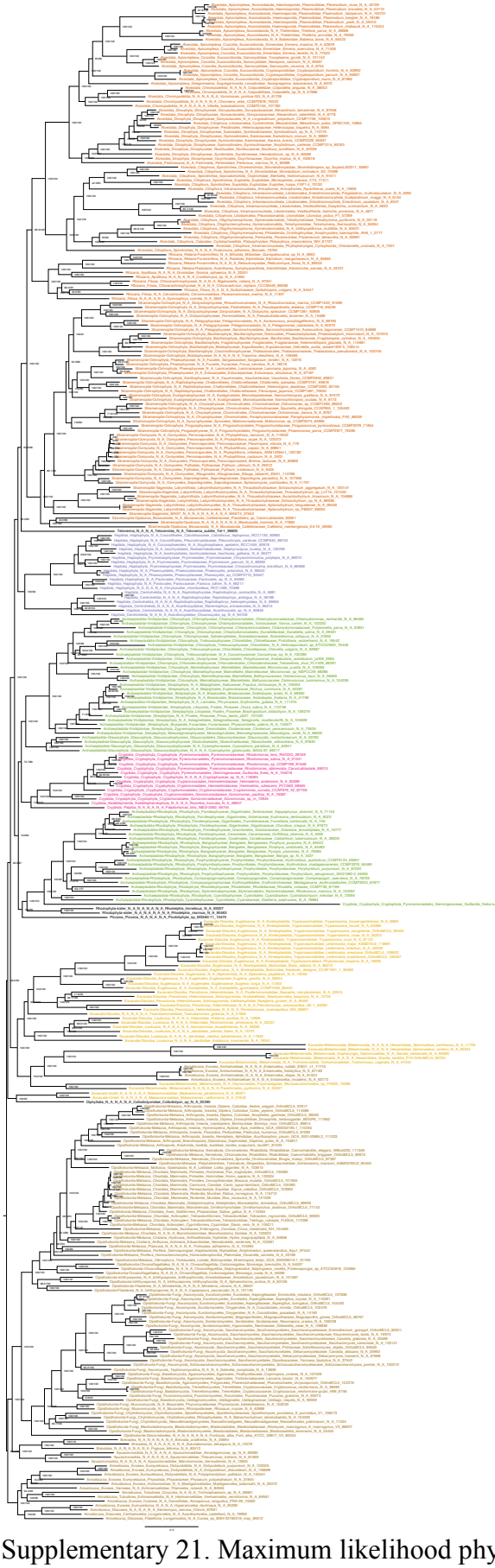
Supplementary Fig. 18. Internode certainty (IC) scores plotted against branch lengths.



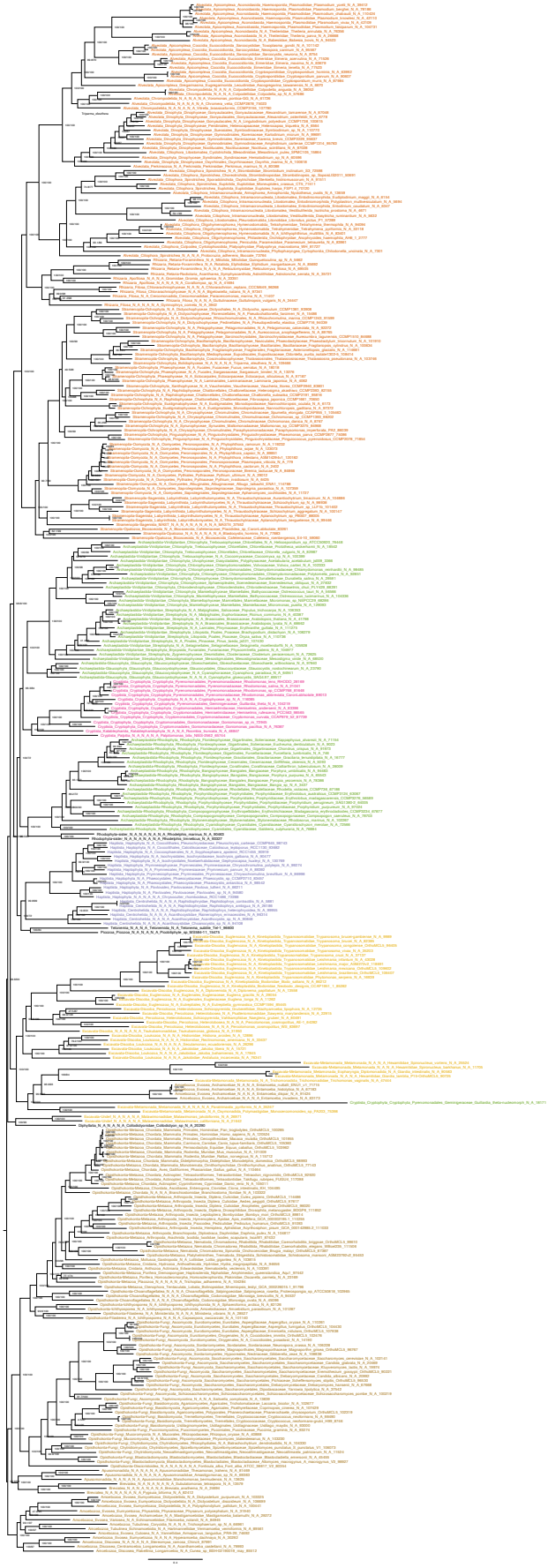
Supplementary Fig. 19. Comparison of performance of three alignment-trimming algorithms: untrimmed (UNTRIM), Divvier partial (DIVPART), or block-trimming (BMGE). 311 gene trees were inferred by maximum likelihood under BIC-selected models on the combined datasets contain up to 344 taxa. (a) Normalized Robinson-Foulds distances (nRF) between gene trees and the corresponding concatenated trees (b) Proportion of highly supported congruent (nHSG), incongruent (nHSB), and lowly supported congruent (nLSG) and incongruent (nLSB) branches. Congruent bipartitions are those present in the corresponding concatenated maximum likelihood trees and high support is considered when SH-aLRT >0.85. (c) nRF distances for short and long gene alignments (>400 aligned amino acids in COMB-UNTRIM). (d) Gene alignment lengths (measured from COMB-UNTRIM alignments) as a function of the best filtering method (producing the lowest nRF). (e) Number of taxa in alignments as a function of the best filtering method (lowest nRF).



Supplementary 20. Maximum likelihood phylogeny of the COMB-BMGE dataset under LG+I+F+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

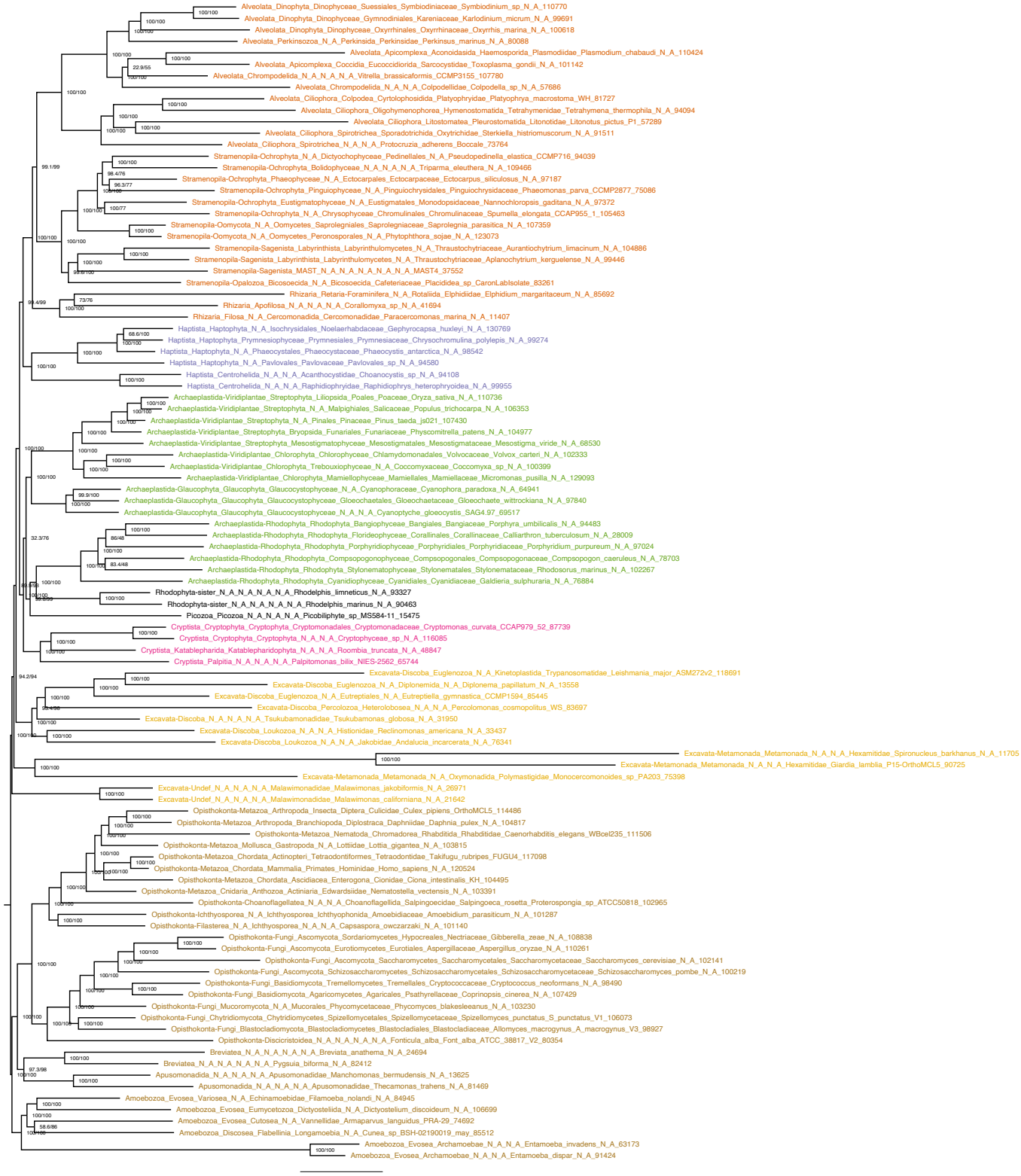


Supplementary 21. Maximum likelihood phylogeny of the COMB-DIVPART dataset under LG+I+Γ. Numbers at nodes are UFBoot / SH-like aLRT supports.



Supplementary 22. Maximum likelihood phylogeny of the COMB-UNTRIM dataset under LG+I+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

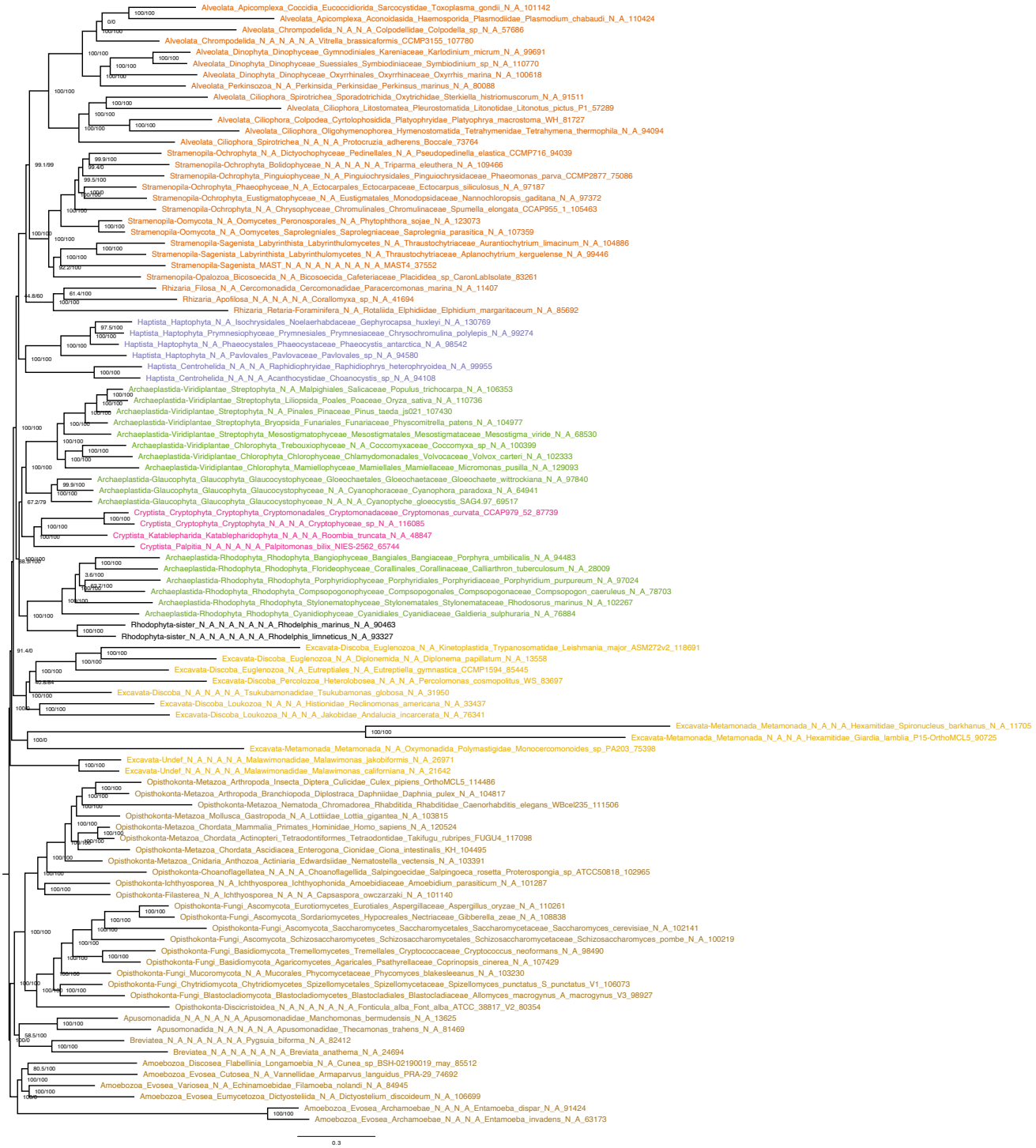
Supplementary Fig. 23. Maximum likelihood phylogeny of the COMB-BMGE-101 dataset under LG+C60+I+F+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.



Supplementary Fig. 24. Maximum likelihood phylogeny of the COMB-DIVPART-101 dataset under LG+C60+I+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

Supplementary Fig. 25. Maximum likelihood phylogeny of the COMB-UNTRIM-101 dataset under LG+C60+I+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

Supplementary Fig. 26. Bayesian phylogeny of the COMB-BMGE-101 dataset under CAT-GTR (majority-rule consensus of three MCMC chains). Numbers at nodes are Bayesian posterior probabilities.

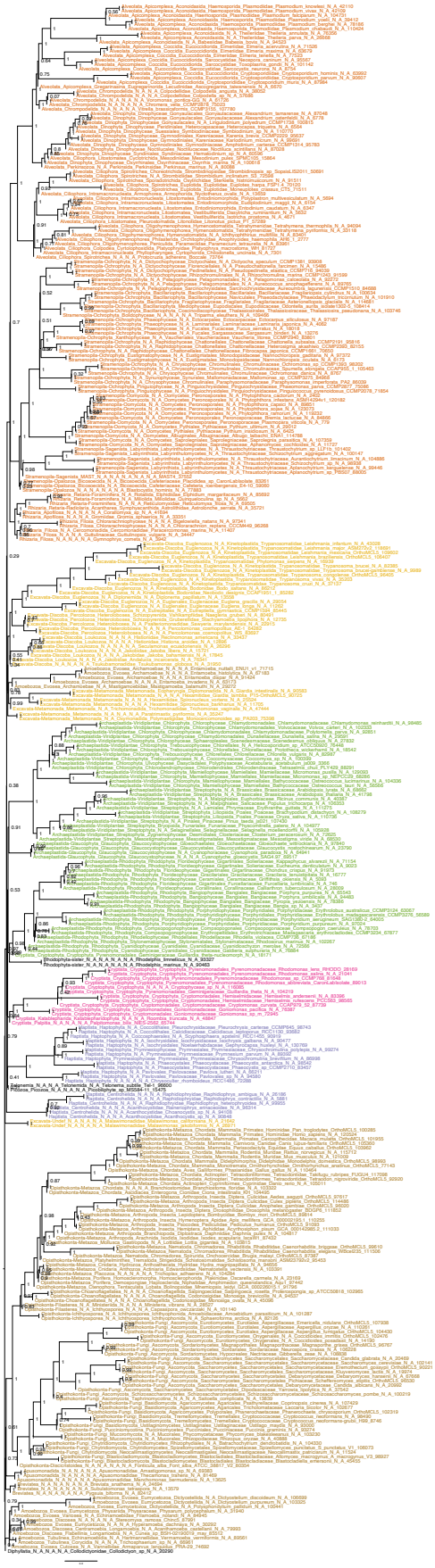


Supplementary Fig. 27. Maximum likelihood phylogeny of the COMB-BMGE-101 dataset without Picozoa under LG+C60+I+F+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

Supplementary Fig. 28. Maximum likelihood phylogeny of the COMB-DIVPART-101 dataset without Picozoa under LG+C60+I+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

Supplementary Fig. 29. Maximum likelihood phylogeny of the COMB-UNTRIM-101 dataset without Picozoa under LG+C60+I+Γ4. Numbers at nodes are UFBoot / SH-like aLRT supports.

Supplementary Fig. 30. Summary coalescent analysis of the COMB-BMGE dataset. Numbers at nodes are local posterior probabilities.



Supplementary Fig. 31. Summary coalescent analysis of the COMB-DIVPART dataset. Numbers at nodes are local posterior probabilities.

Supplementary Fig. 32. Summary coalescent analysis of the COMB-UNTRIM dataset. Numbers at nodes are local posterior probabilities.