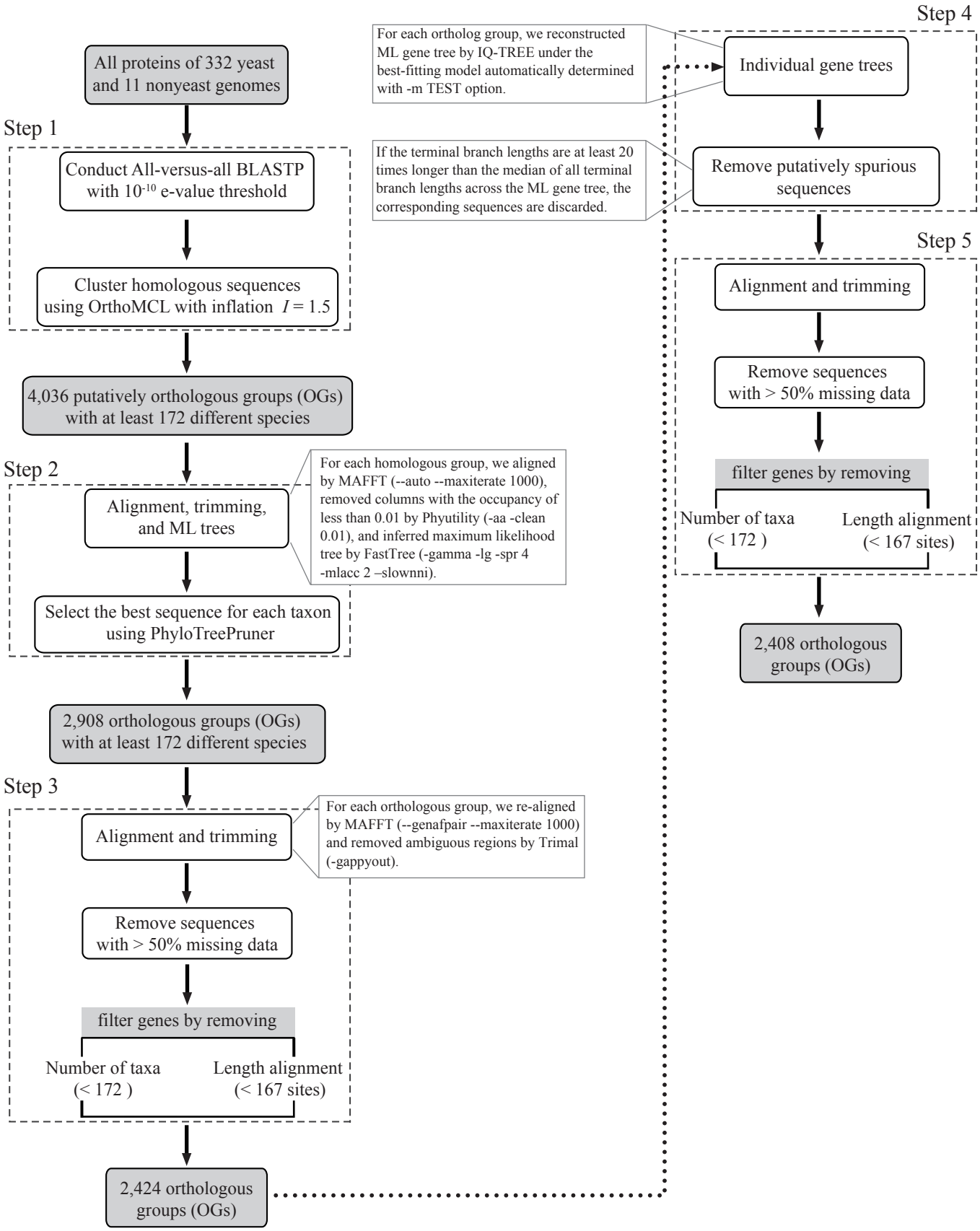
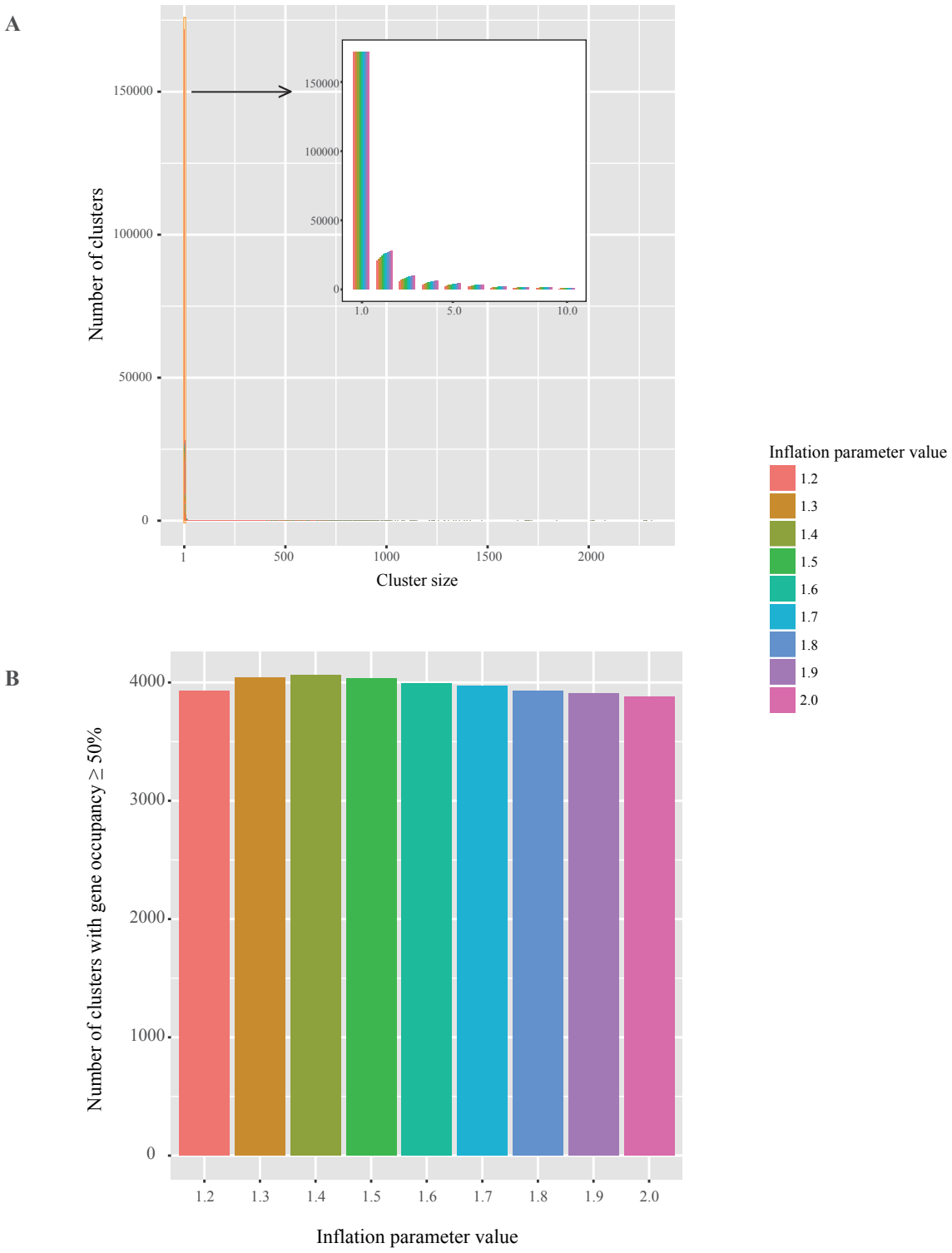
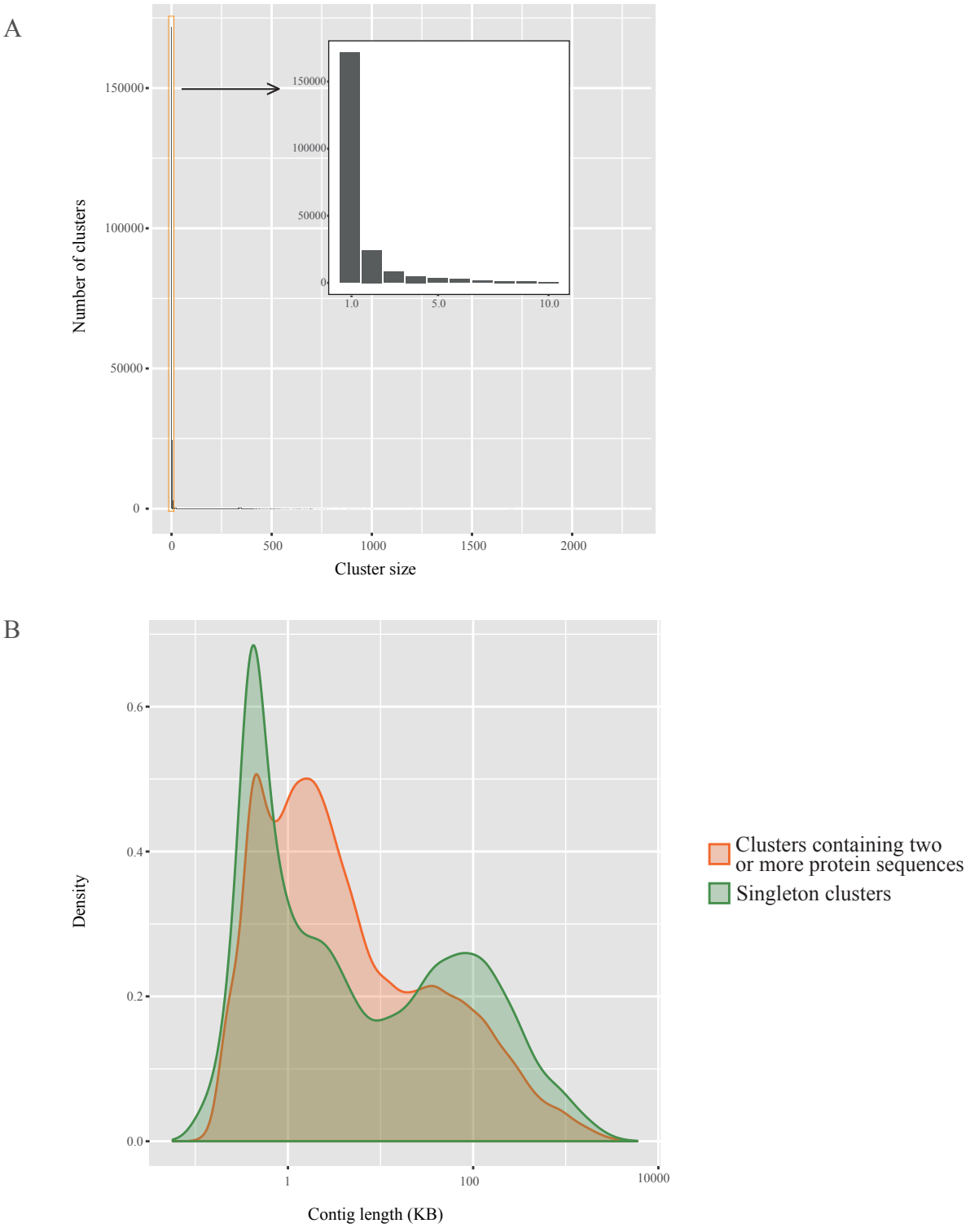


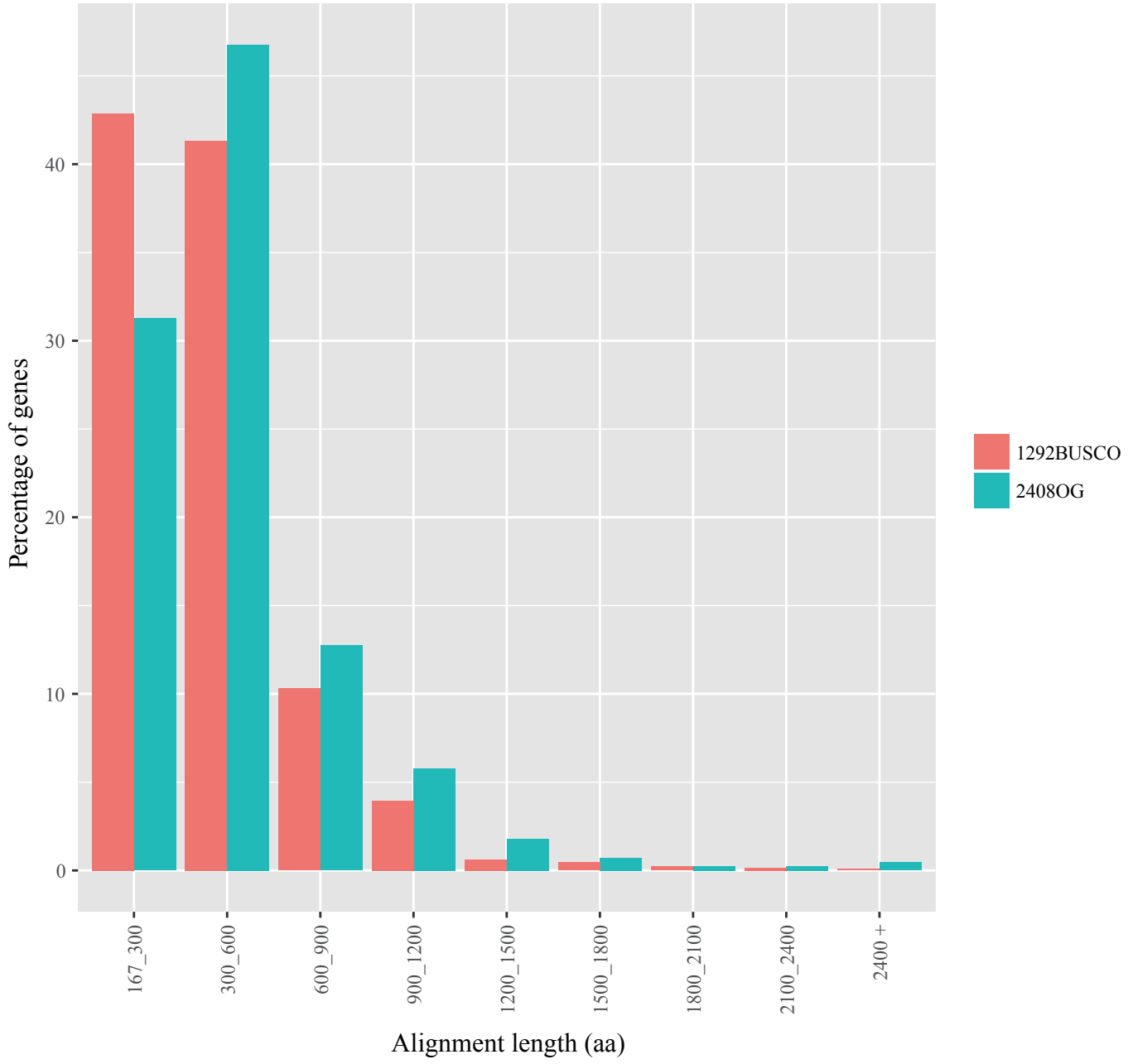
Additional Figure 2. The workflow used for the construction of the 2408 orthologous group (OG) data matrix. A detailed description of the analyses performed in each step of the workflow is provided in the "Phylogenomic data matrix construction" section of the Supplementary Methods.



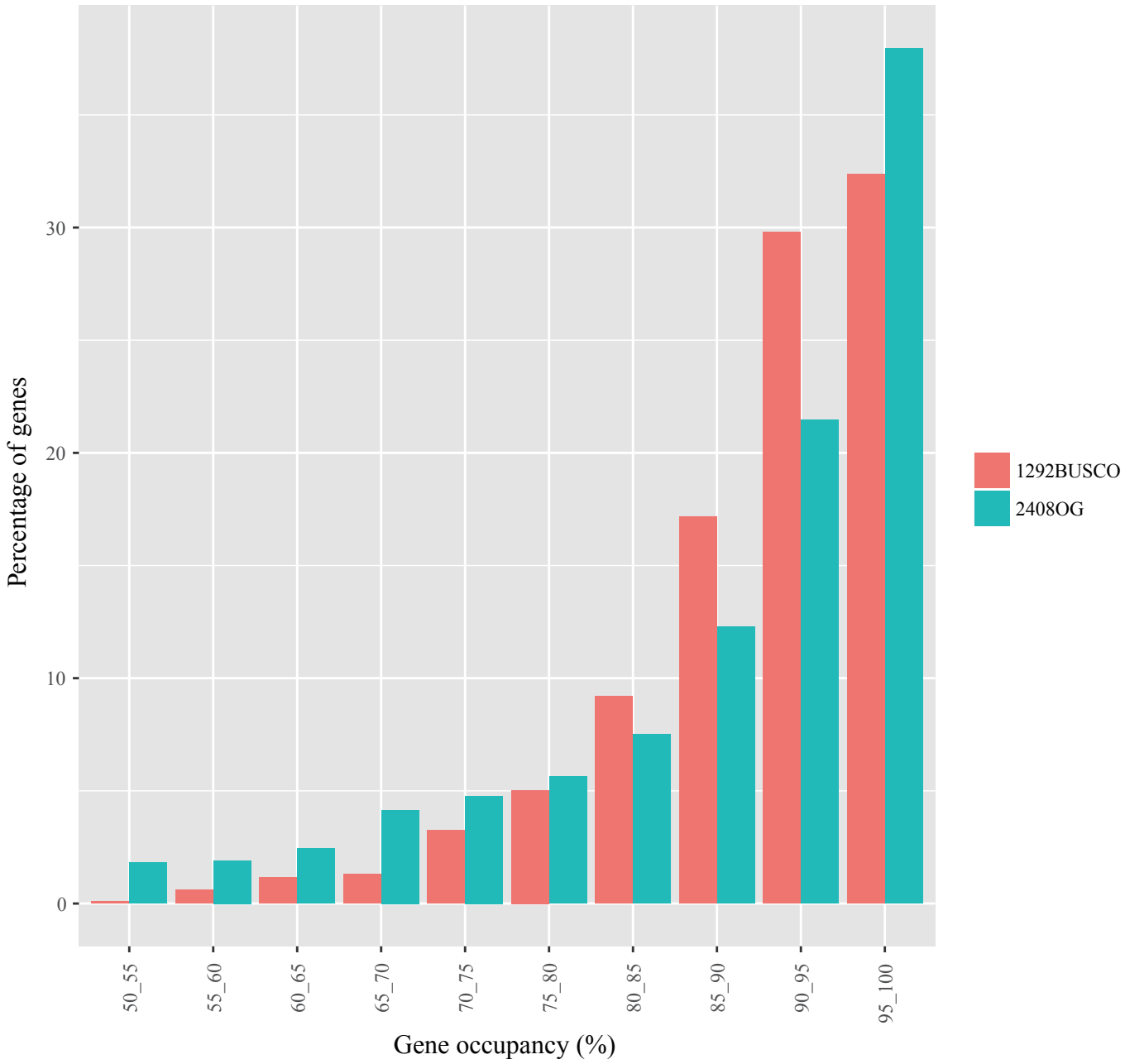
Additional Figure 3. The effect of inflation parameter value on the identification of clusters of homologous genes using the OrthoMCL program. (A) To compare the distributions of cluster sizes (i.e., the distributions of the number of protein sequences assigned to each cluster) produced by the OrthoMCL program, we used a range of inflation parameter values from 1.2 to 2.0 (with a step of 0.1). The distributions of cluster sizes from 1 to 10 are displayed in the inset. The distributions of cluster sizes were very similar across inflation parameter values and resulted in very high proportions of clusters containing one or very few genes. These very high proportions stem from (i) the conservative BLASTP e-value cutoff (10-10) that we used, and (ii) comparing genomes that exhibit substantial levels of genetic divergence (as we show in Figure 1, budding yeasts in the subphylum exhibit a range of divergences on par to those observed among animal genomes). (B) The effect of inflation parameter values on numbers of cluster sizes with gene occupancy ≥ 50%, that is those clusters that contained one or more protein sequences from at least half (≥ 172) of the 343 genomes (332 budding yeasts and 11outgroups). These distributions were very similar across inflation parameter values. Given these results and a previous study showing that an inflation parameter value of 1.5 was optimal for ortholog identification in budding yeasts (Salichos and Rokas, 2011), we decided to use the OrthoMCL clusters obtained when the inflation parameter value was set to 1.5 (in green) for construction of the 2408OG phylogenomic data matrix.



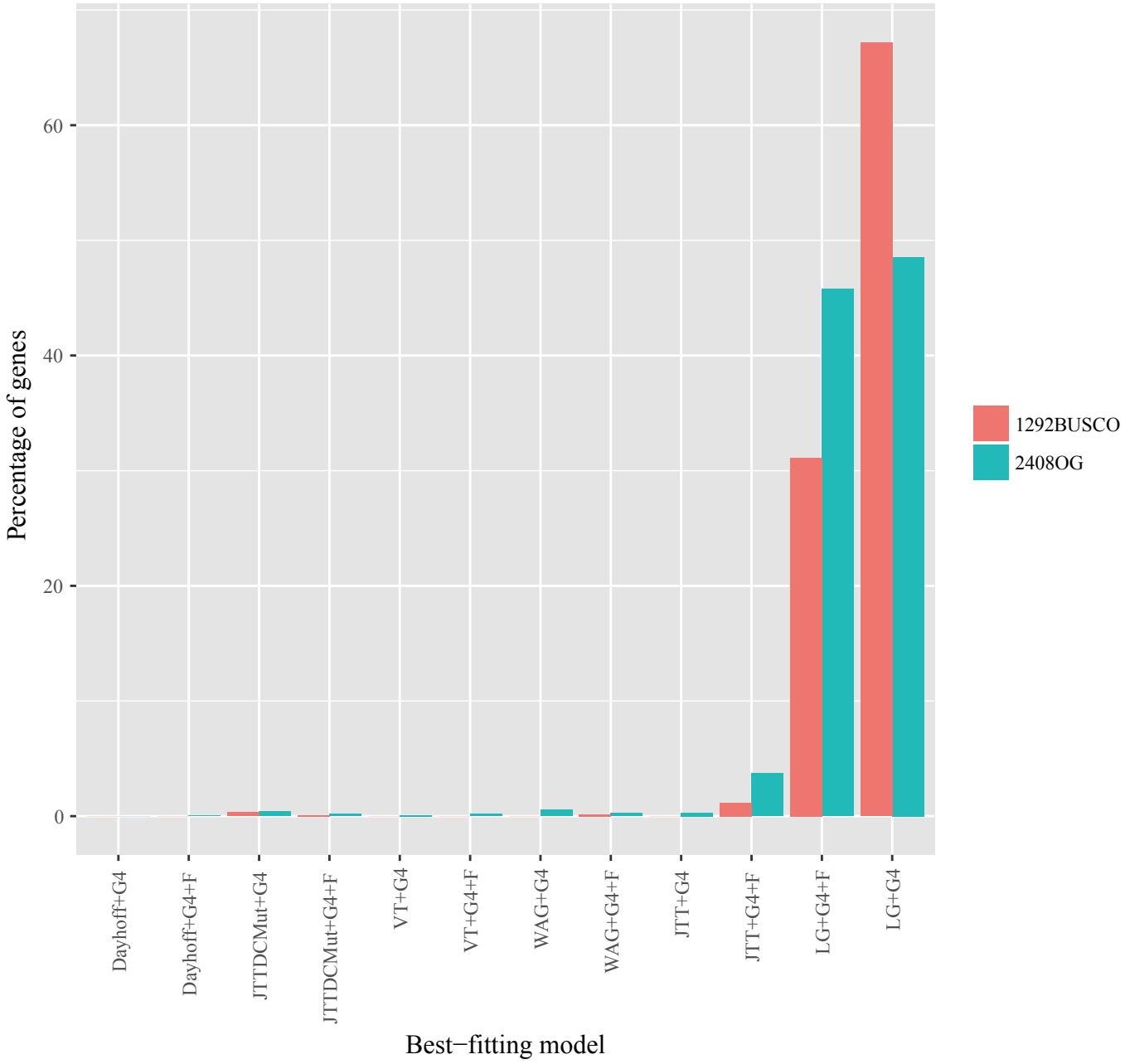
Additional Figure 4. Distribution of sizes of clusters of homologous proteins identified using OrthoMCL with an inflation parameter value of 1.5 (panel A) and distribution of the lengths of the genomic contigs that their proteins reside in (panel B). (A) The distribution of cluster sizes of the 233,478 clusters identified using the OrthoMCL program with an inflation parameter value of 1.5. The distributions of cluster sizes from 1 to 10 are displayed in the inset. 171,715 clusters contain a single protein (singleton clusters) and 61,763 clusters contain two or more proteins. (B) The distribution of the lengths of genomic contigs that contain proteins present in the 171,715 singleton clusters compared to the distribution of the lengths of genomic contigs present in the 61,763 non-singleton clusters.



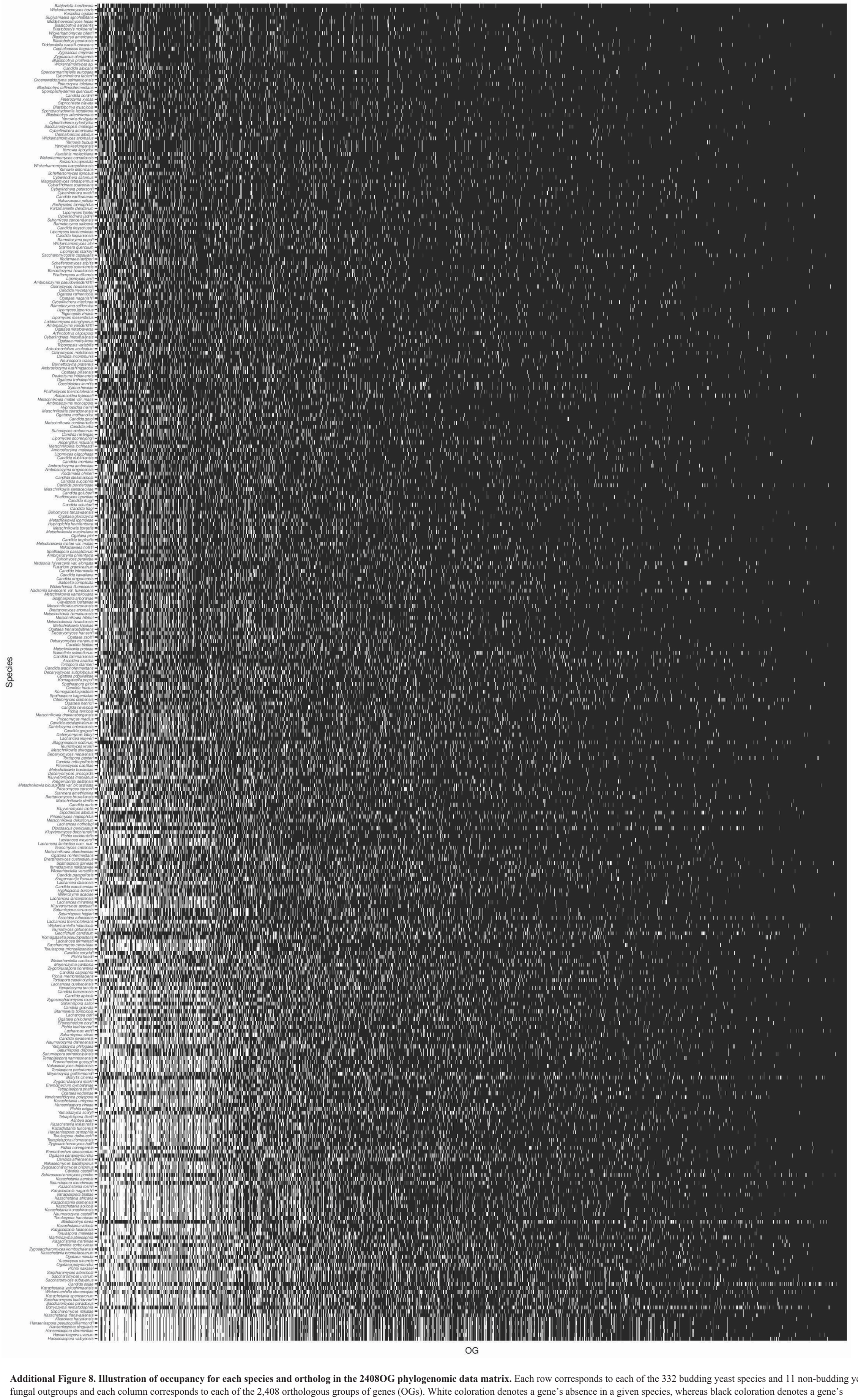
Additional Figure 5. Distribution of gene alignment lengths for the 2408OG and 1292BUSCO phylogenomic data matrices. The sequence alignment length (aa, amino acid residues) of each gene in each data matrix was measured after filtering out ambiguously aligned positions.



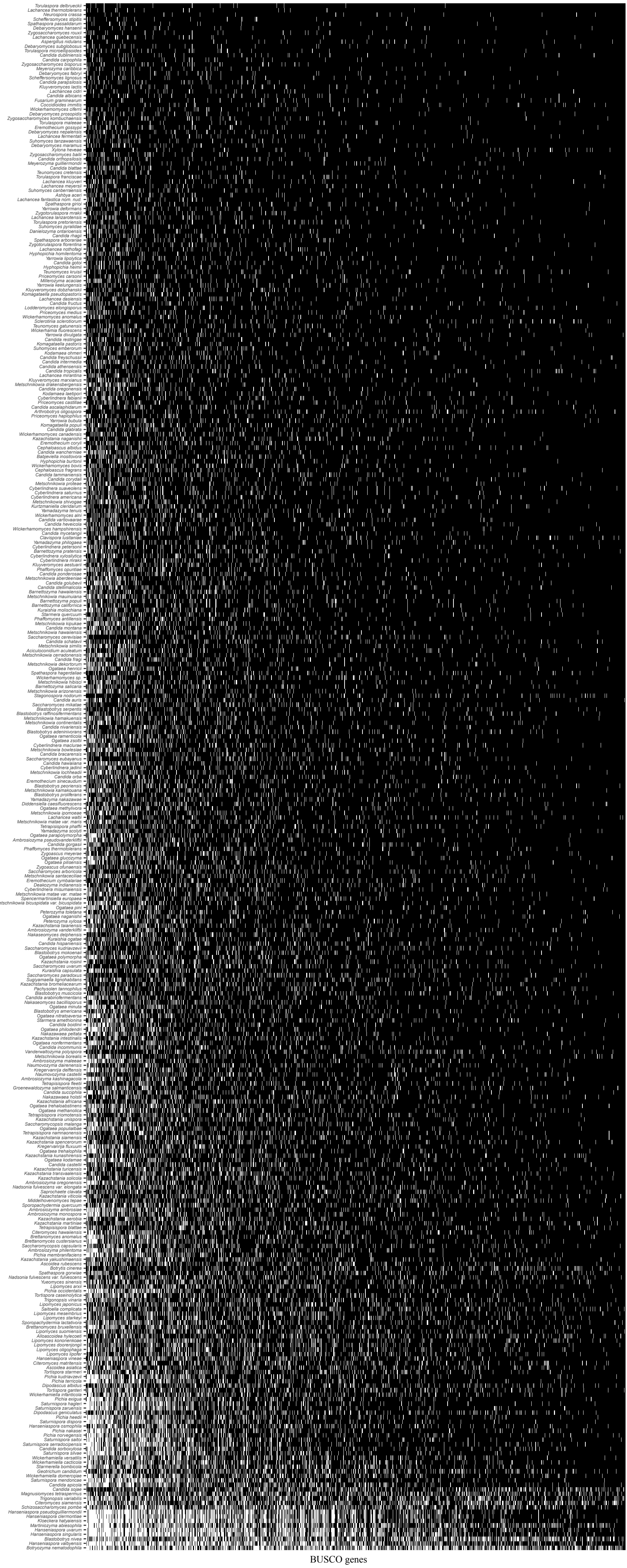
Additional Figure 6. Distribution of gene occupancy for the 2408OG and 1292BUSCO phylogenomic data matrices. Gene occupancy (also known as taxon coverage) was calculated as (number of taxa with sequences available in the trimmed alignment) / 343 (332 yeast taxa + 11 outgroups).



Additional Figure 7. Distribution of the best-fitting amino acid substitution models for the 2408OG and 1292BUSCO data matrices. The best-fitting model for each gene in each data matrix was determined using IQ-TREE multicore version 1.5.1 under the Bayesian information criterion (BIC).



Additional Figure 8. Illustration of occupancy for each species and ortholog in the 2408OG phylogenomic data matrix. Each row corresponds to each of the 332 budding yeast species and 11 non-budding yeast fungal outgroups and each column corresponds to each of the 2,408 orthologous groups of genes (OGs). White coloration denotes a gene's absence in a given species, whereas black coloration denotes a gene's presence in a given species. OGs exhibiting the highest occupancy across species are toward the right of the graph, and species exhibiting the highest occupancy across OGs are toward the top of the graph. The 2408OG data matrix contains 729,289 genes (~36% of the total number of genes).



BUSCO genes

Additional Figure 9. Illustration of occupancy for each species and ortholog in the 1292BUSCO phylogenomic data matrix. Each row corresponds to each of the 332 budding yeast species and 11 non-budding yeast fungal outgroups and each column corresponds to each of the 1,292 BUSCO genes. White coloration denotes a gene's absence in a given species, whereas black coloration denotes a gene's presence in a given species. BUSCO genes exhibiting the highest occupancy across species are toward the right of the graph, and species exhibiting the highest occupancy across BUSCO genes are toward the top of the graph. The 1292BUSCO data matrix contains 400,449 genes (~20% of the total number of genes).

Additional Figure 10. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the concatenation-based analysis of 2,408 genes in the 2408OG data matrix under a single LG + GAMMA. The ML phylogeny (*lnL*=-286,053,422.906) was reconstructed using the 2408OG data matrix (1,162,805 sites) under a single LG + GAMMA substitution model using IQ-TREE multicore version 1.5.1. Branch support values near internodes correspond to bootstrap support (below) and internode certainty (above), respectively. Only bootstrap support values smaller than 90% and internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 10) and the coalescence-based phylogeny (Additional Figure12). The concatenation-based ML phylogenies under a single partition (Additional Figure 10) and under gene-based partitioning (additional Figure 11) are topologically identical, with the only exception being the bipartition defined by internal branch I246, which is shown in Additional Figures 18 and 19. Note that the ML phylogram with branch lengths corresponding to amino acid substitutions / site is shown at the bottom left.

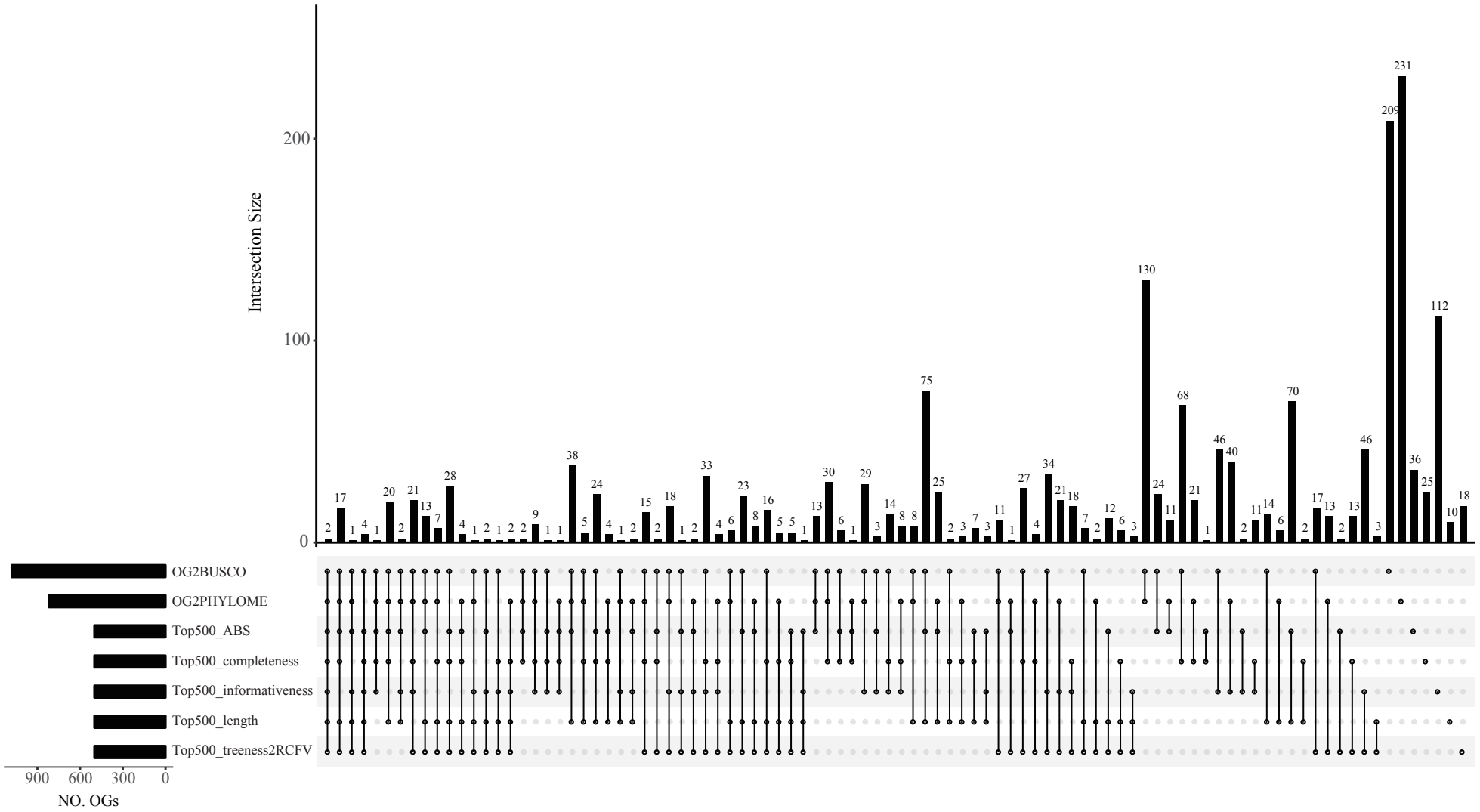
Additional Figure 11. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the concatenation-based analysis of 2,408 genes in the 2408OG data matrix under the gene-based partitioning. The ML phylogeny (*lnL*= -285,074,794.532) was reconstructed using the 2408OG data matrix (1,162,805 sites) under gene-based partitioning using IQ-TREE multicore version 1.5.1. Branch support values near internodes correspond to bootstrap support (below) and internode certainty (above), respectively. Only bootstrap support values smaller than 90% and internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 11) and the coalescence-based phylogeny (Additional Figure 12). The concatenation-based ML phylogenies under a single partition (Additional Figure 10) and under gene-based partitioning (Additional Figure 11) are topologically identical, with the only exception being the bipartition defined by internal branch I246, which is shown in Additional Figures 18 and 19. Note that the ML phylogram with branch lengths corresponding to amino acid substitutions / site is shown at the bottom left.

Additional Figure 12. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the coalescence-based analysis of 2,408 genes in the 2408OG data matrix. The coalescence-based phylogenetic tree was reconstructed using ASTRAL-II version 4.10.2. Branch support values near internodes correspond to local posterior probability (below) and internode certainty (above), respectively. Only local posterior probability values smaller than 0.9 and internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 10) and the coalescence-based phylogeny (Additional Figure 12).

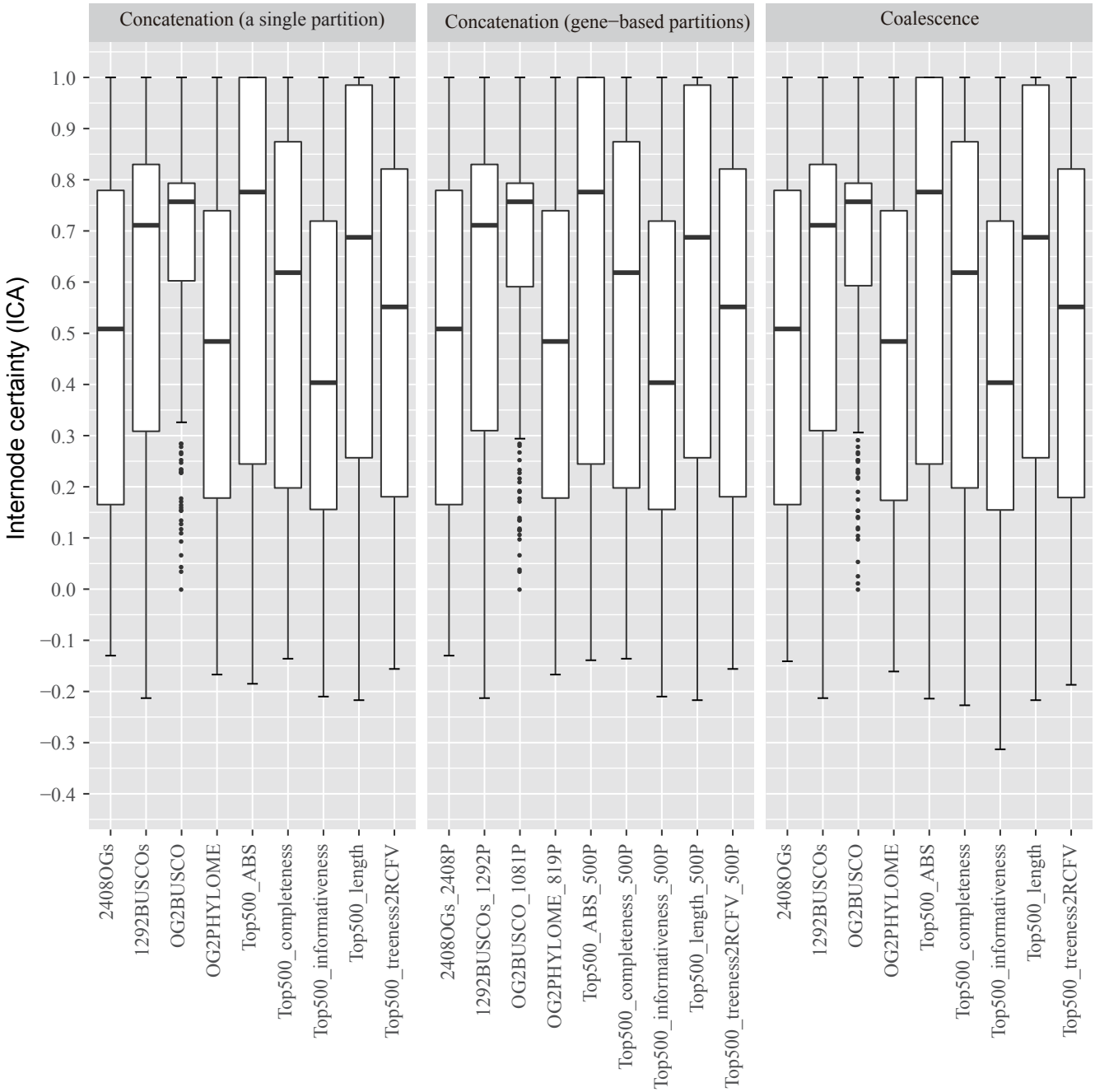
Additional Figure 13. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the concatenation-based analysis of 1,292 genes in the 1292BUSCO data matrix under a single LG + GAMMA. The ML phylogeny (*lnL*=-97,177,023.925) was reconstructed using the 1292BUSCO data matrix (527,069 sites) under a single LG + GAMMA substitution model using IQ-TREE multicore version 1.5.1. Branch support values near internodes correspond to bootstrap support (below) and internode certainty (above), respectively. Only bootstrap support values smaller than 90% and internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 13) and the coalescence-based phylogeny (Additional Figure 15). The concatenation-based ML phylogenies under a single partition (Additional Figure 13) and under gene-based partitioning (Additional Figure 14) are topologically identical, with the only exception being the bipartition defined by internal branch I262, which is shown in Additional Figures 18 and 19. Note that the ML phylogram with branch lengths corresponding to amino acid substitutions / site is shown at the bottom left.

Additional Figure 14. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the concatenation-based analysis of 1,292 genes in the 1292BUSCO data matrix under the gene-based partitioning. The ML phylogeny (*lnL*= -96,950,014.165) was reconstructed using the 1292BUSCO data matrix (527,069 sites) under gene-based partitioning using IQ-TREE multicore version 1.5.1. Branch support values near internodes correspond to bootstrap support (below) and internode certainty (above), respectively. Only bootstrap support values smaller than 90% or internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 14) and the coalescence-based phylogeny (Additional Figure 15). The concatenation-based ML phylogenies under a single partition (Additional Figure 13) and under gene-based partitioning (Additional Figure 14) are topologically identical, with the only exception being the bipartition defined by internal branch I262, which is shown in Additional Figures 18 and 19. Note that the ML phylogram with branch lengths corresponding to amino acid substitutions / site is shown at the bottom left.

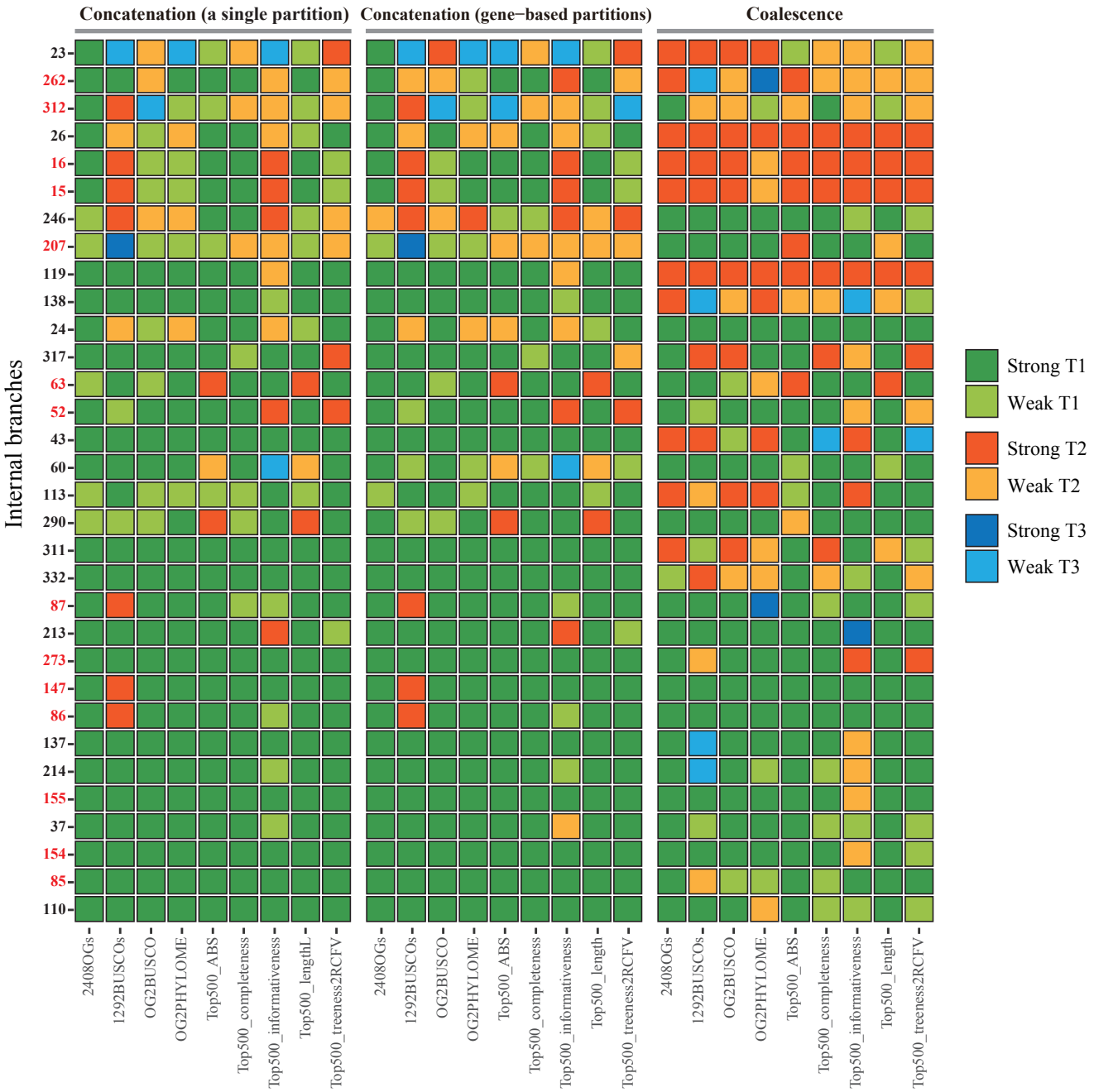
Additional Figure 15. The phylogenetic relationships of 332 budding yeasts and 11 outgroups inferred from the coalescence-based analysis of 1,292 genes in the 1292BUSCO data matrix. The coalescence-based phylogenetic tree was reconstructed using ASTRAL-II version 4.10.2. Branch support values near internodes correspond to local posterior probability (below) and internode certainty (above), respectively. Only local posterior probability values smaller than 0.9 and internode certainty values smaller than 0.1 are shown. Red branches show conflicts between the concatenation-based phylogeny (Additional Figure 13) and the coalescence-based phylogeny (Additional Figure 15).



Additional Figure 16. Relatively few genes are shared across the 7 additional subsampled data matrices. The plot illustrates that the 7 additional data matrices that we created from subsampling of the 2408OG data matrix have relatively few overlapping genes. For example, the first intersect on the left shows that only two genes are shared across all 7 data matrices, whereas the last 7 intersects on the right show the numbers of genes uniquely present in each of the 7 data matrices. The intersecting plot was generated using the UpSetR R package (https://cran.r-project.org/web/packages/UpSetR/).

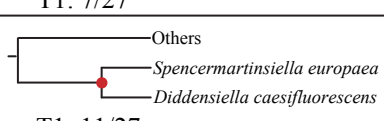
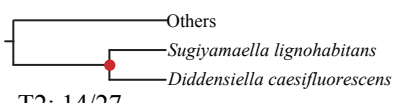
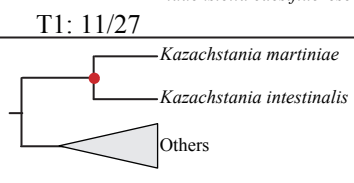
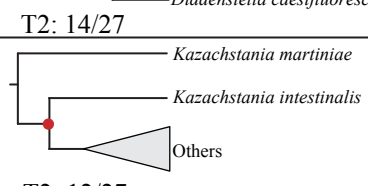
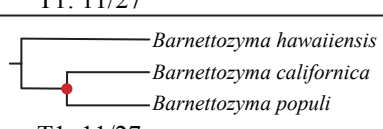
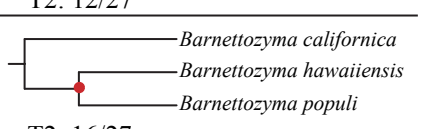
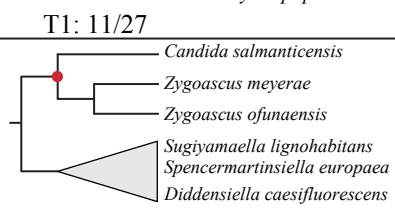
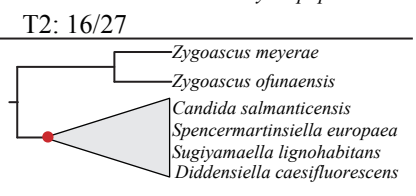
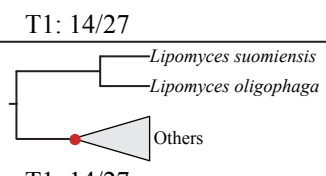
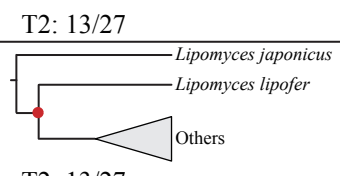
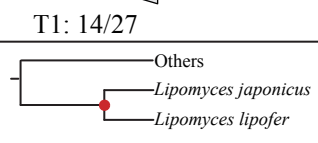
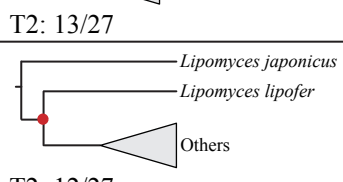
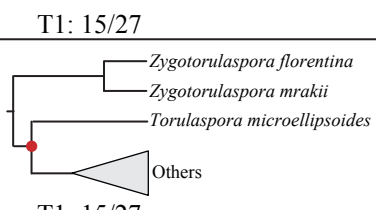
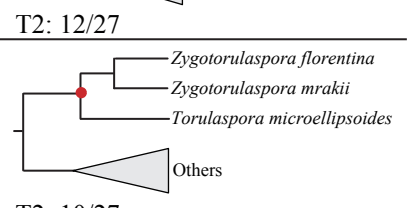
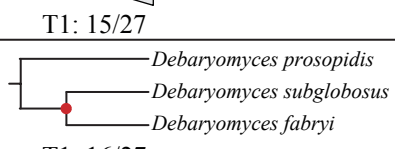
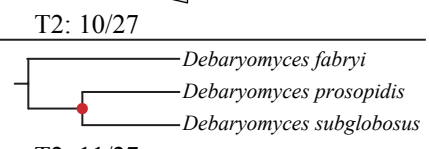
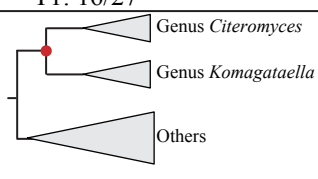
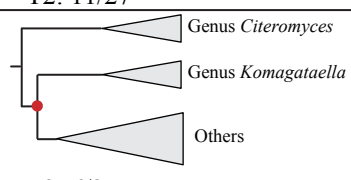
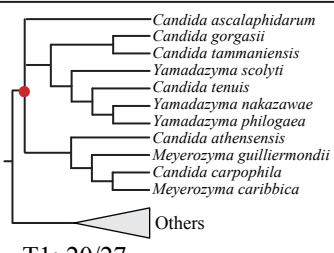
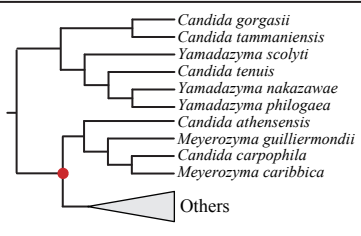
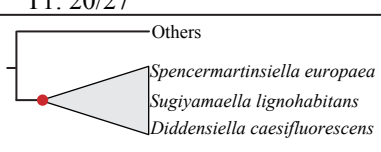
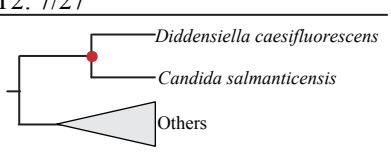
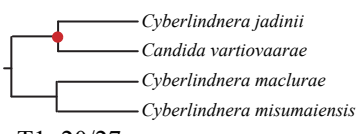
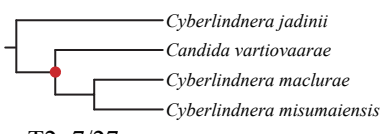
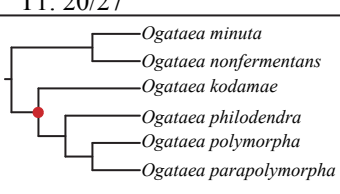
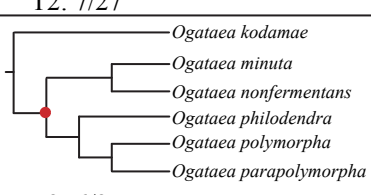
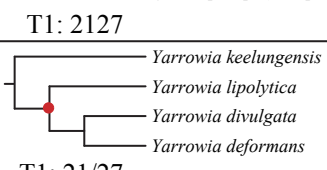
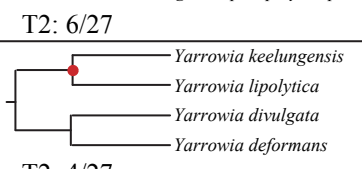
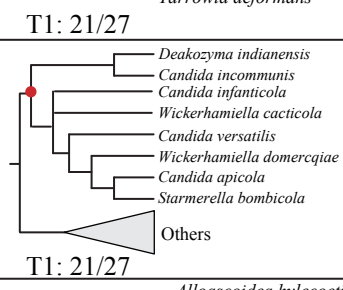
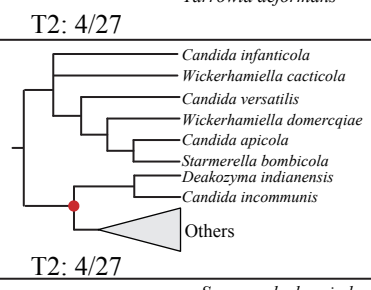
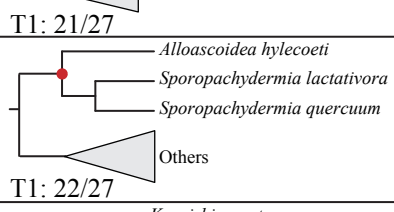
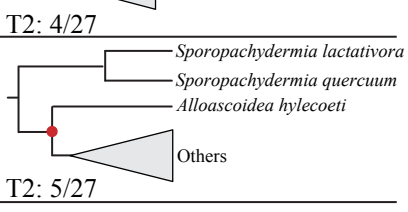
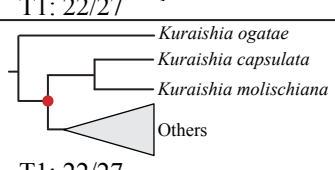
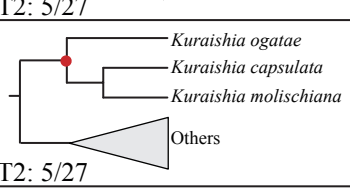
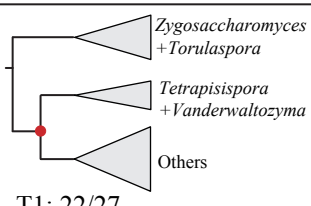
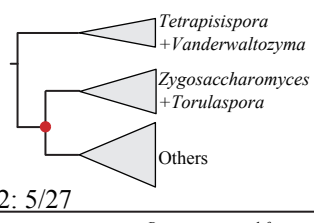
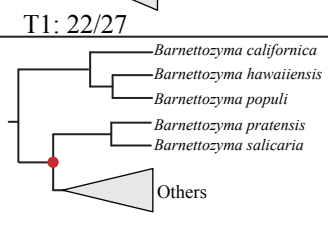
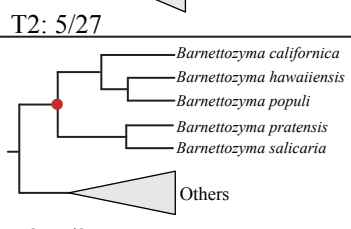
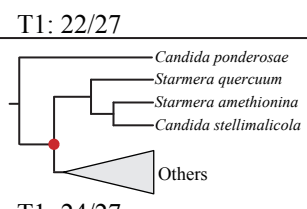
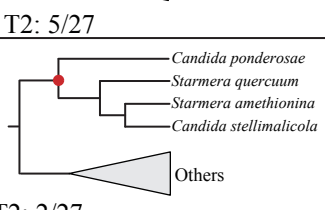


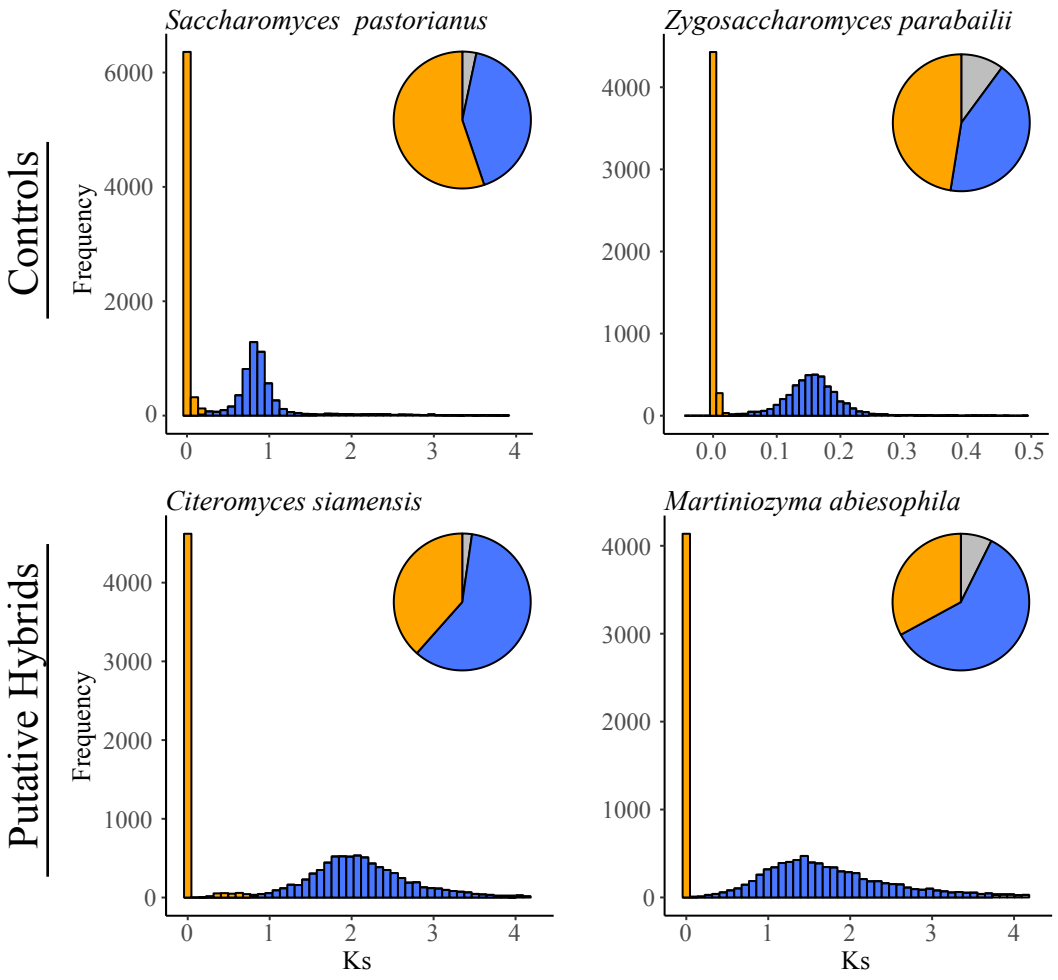
Additional Figure 17. Boxplot of internode certainty scores in the each of 27 phylogenies. The 27 phylogenies were reconstructed from analyses of 9 different data matrices (2408OG, 1292BUSCO, and the 7 data matrices constructed from subsamples of the 2408OG data matrix) using three different approaches (concatenation under a single partition, concatenation under gene-based partitioning, and coalescence). For each phylogeny, internode certainty (ICA) scores for all 340 internal branches were calculated from the set of individual ML gene trees using RAxML with the option '-f i'. Rectangles in the boxplot denote 1st and 3rd quartiles, horizontal thick bars represent mean ICA values, and dots correspond to outliers whose ICA scores are greater than two standard deviations from the mean.



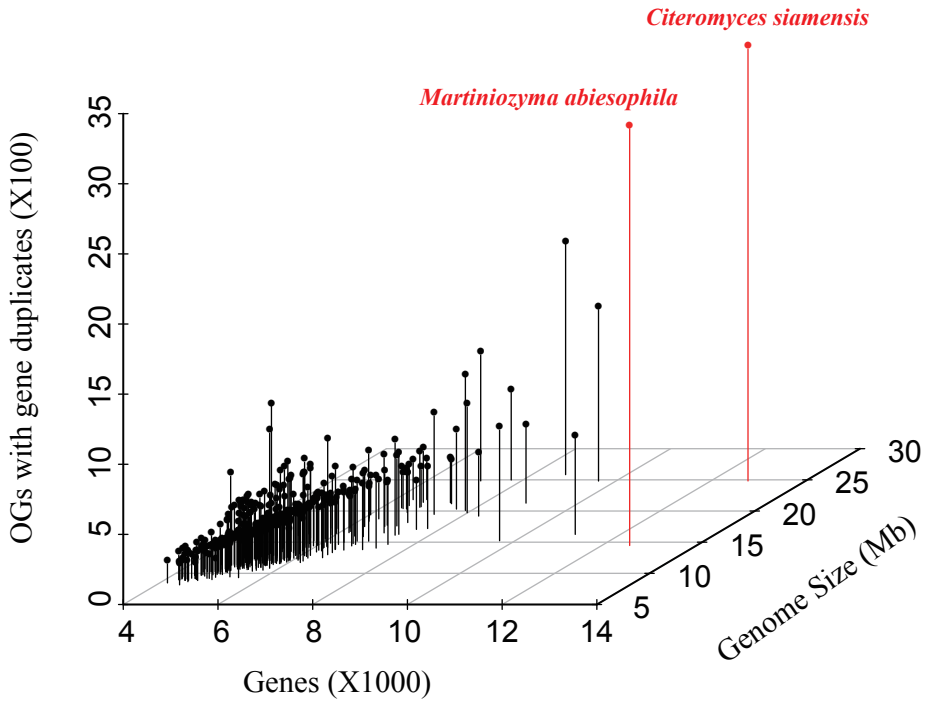
Additional Figure 18. Summary of topological support for the 32 internal branches that are incongruent across the 27 phylogenies of the 332 Saccharomycotina yeasts. The 27 phylogenies were reconstructed from analyses of 9 different data matrices (2408OG, 1292BUSCO, and the 7 data matrices constructed from subsamples of the 2408OG data matrix using three different approaches (concatenation under a single partition, concatenation under gene-based partitioning, and coalescence). Internal branches with bootstrap support values ≥ 90 or local probability values ≥ 0.9 were considered to be strongly supported (darker colors), while all others were considered to be weakly supported (lighter colors). Internal branches (bipartitions) present in the ML phylogeny inferred from analysis of the 2408OG data matrix under concatenation and a single LG+G4 model of amino acid sequence evolution (Figure 2) were considered to be the reference bipartitions and were labeled as T1 (Tree 1). The bipartition showing the highest degree of conflict with T1 (among the 27 phylogenies) was labeled as T2 (Tree 2), and the bipartition showing the next highest degree of conflict as T3 (Tree 3). Internal branch names on the Y axis correspond to the branches shown in Figure 2 and in Additional Figure 19. The 14 internal branch names shown in red font correspond to conflicts observed within genera, whereas the 18 internal branch names shown in black font correspond to the conflicts observed between genera or higher taxonomic ranks. Detailed topologies of each of these 32 internal branches are shown in Additional Figure 19.

Additional Figure 19. Detailed topologies of alternative hypotheses for the 32 internal branches that show conflict across the 27 phylogenies of 332 Saccharomycotina yeasts. For each of the 32 internal branches that exhibit conflict, the T1 topology (found in the phylogeny shown in Figure 2) and the T2 topology (i.e., the topology with the highest degree of conflict among the 27 phylogenies) are shown. Internal branch names correspond to branches shown in Figure 2 and Additional Figures 18 and 25.

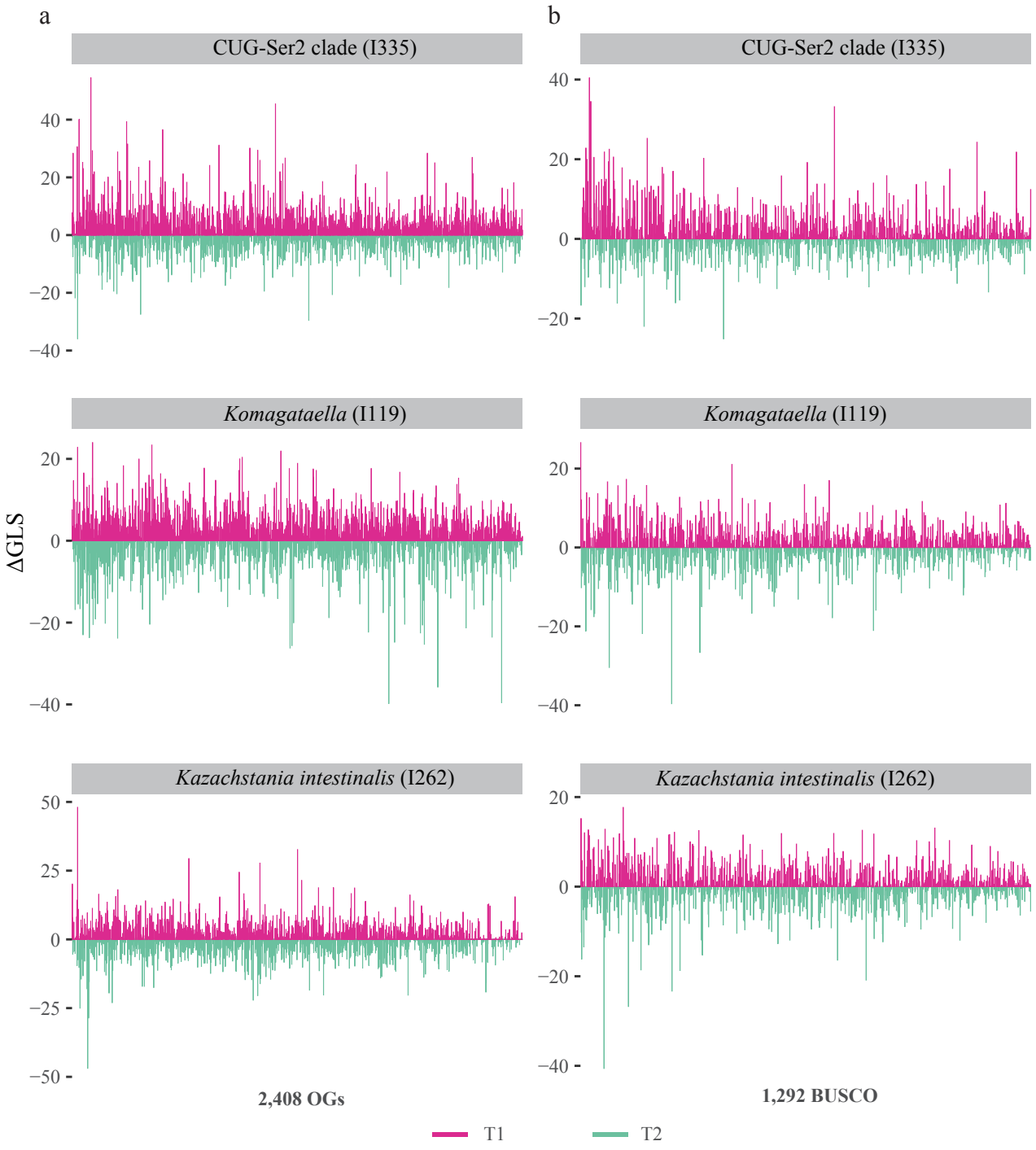
Internal branch	T1 topology	The most conflicting topology
I23	T1: 7/27	T2: 13/27
		
I262	T1: 11/27	T2: 14/27
		
I312	T1: 11/27	T2: 12/27
		
I26	T1: 11/27	T2: 16/27
		
I16	T1: 14/27	T2: 13/27
		
I15	T1: 14/27	T2: 13/27
		
I246	T1: 15/27	T2: 12/27
		
I207	T1: 15/27	T2: 10/27
		
I119	T1: 16/27	T2: 11/27
		
I138	T1: 19/27	T2: 6/27
		
I24	T1: 20/27	T2: 7/27
		
I317	T1: 20/27	T2: 7/27
		
I63	T1: 20/27	T2: 7/27
		
I52	T1: 2127	T2: 6/27
		
I43	T1: 21/27	T2: 4/27
		
I60	T1: 21/27	T2: 4/27
		
I113	T1: 22/27	T2: 5/27
		
I290	T1: 22/27	T2: 5/27
		
I311	T1: 22/27	T2: 5/27
		
I332	T1: 22/27	T2: 5/27
		
I87	T1: 24/27	T2: 2/27
I213	T1: 24/27	T2: 2/27
I273	T1: 24/27	T2: 3/27
I147	T1: 25/27	T2: 2/27
I86	T1: 25/27	T2: 2/27
I137	T1: 25/27	T2: 1/27
I214	T1: 24/27	T2: 1/27
I155	T1: 26/27	T2: 1/27
I37	T1: 26/27	T2: 1/27
I154	T1: 26/27	T2: 1/27
I85	T1: 26/27	T2: 1/27
I110	T1: 26/27	T2: 1/27



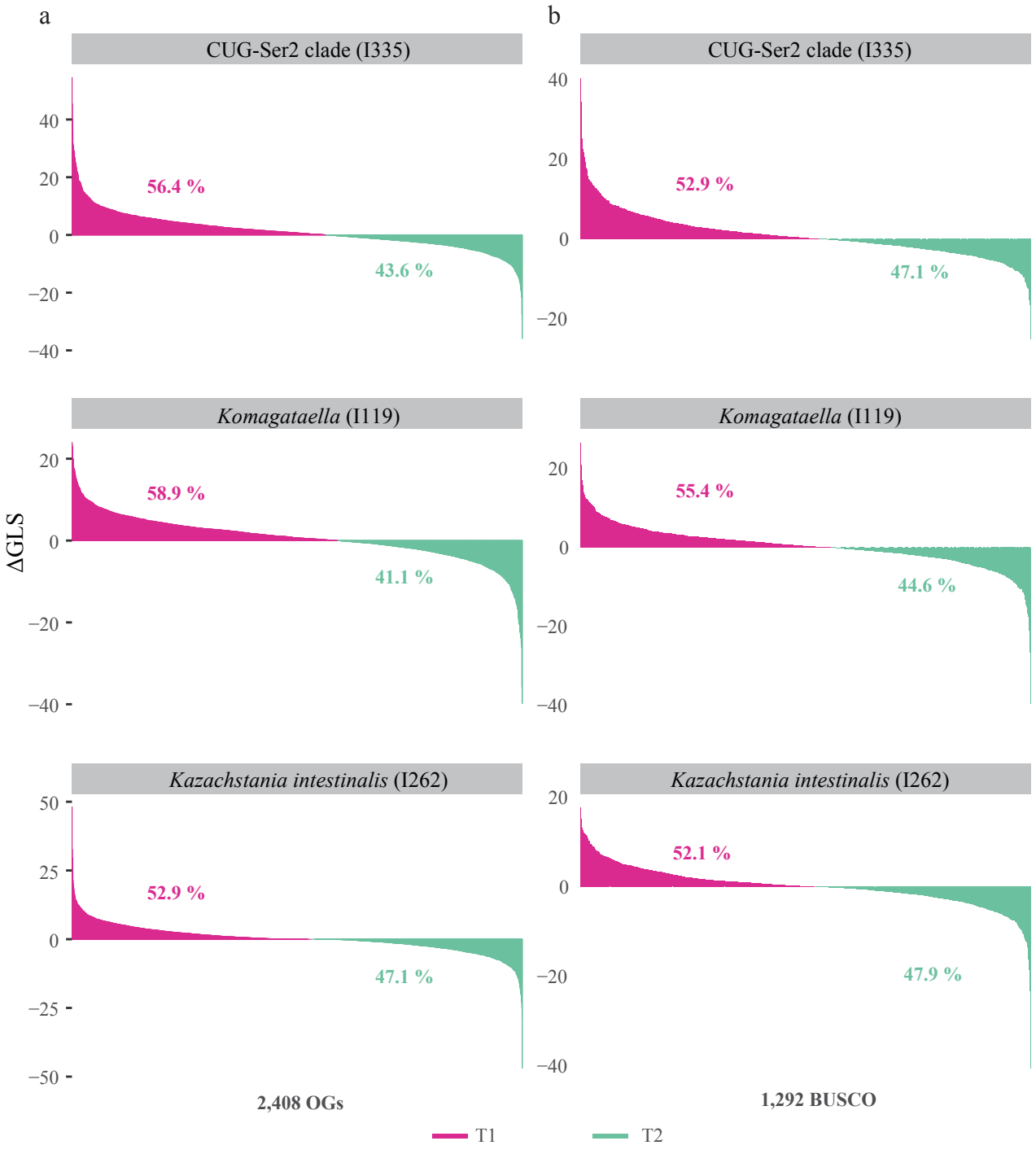
Additional Figure 20. Bimodal distributions of the number of synonymous substitutions per synonymous site (*Ks*) suggest that the genomes of *Citeromyces siamensis* and *Martiniozyma abiesophila* are of hybrid origin. The two panels in the top row depict the distributions of number of synonymous substitutions per synonymous site (*Ks*) for *Saccharomyces pastorianus* (left) and *Zygosaccharomyces parabailii* (right), two previously reported hybrids that we used as controls in the present analysis; both show bimodal distributions of *Ks*, reflecting the fact that some genes in the hybrid genome are most closely related to one parental species (orange bars) and some to the other more divergent parent (blue bars). Grey bars correspond to genes that could not be reliably assigned to either parental species. The two panels in the bottom row depict the distributions of *Ks* for *Citeromyces siamensis* and *Martiniozyma abiesophila*, two of our newly sequenced genomes. Both genomes show bimodal *Ks* distributions similar to those observed in known hybrids, suggesting that they are also of hybrid origin. *Ks* was calculated between all pairs of protein-coding homologs from the species of interest and its closest relative on the phylogeny shown in Figure 2.



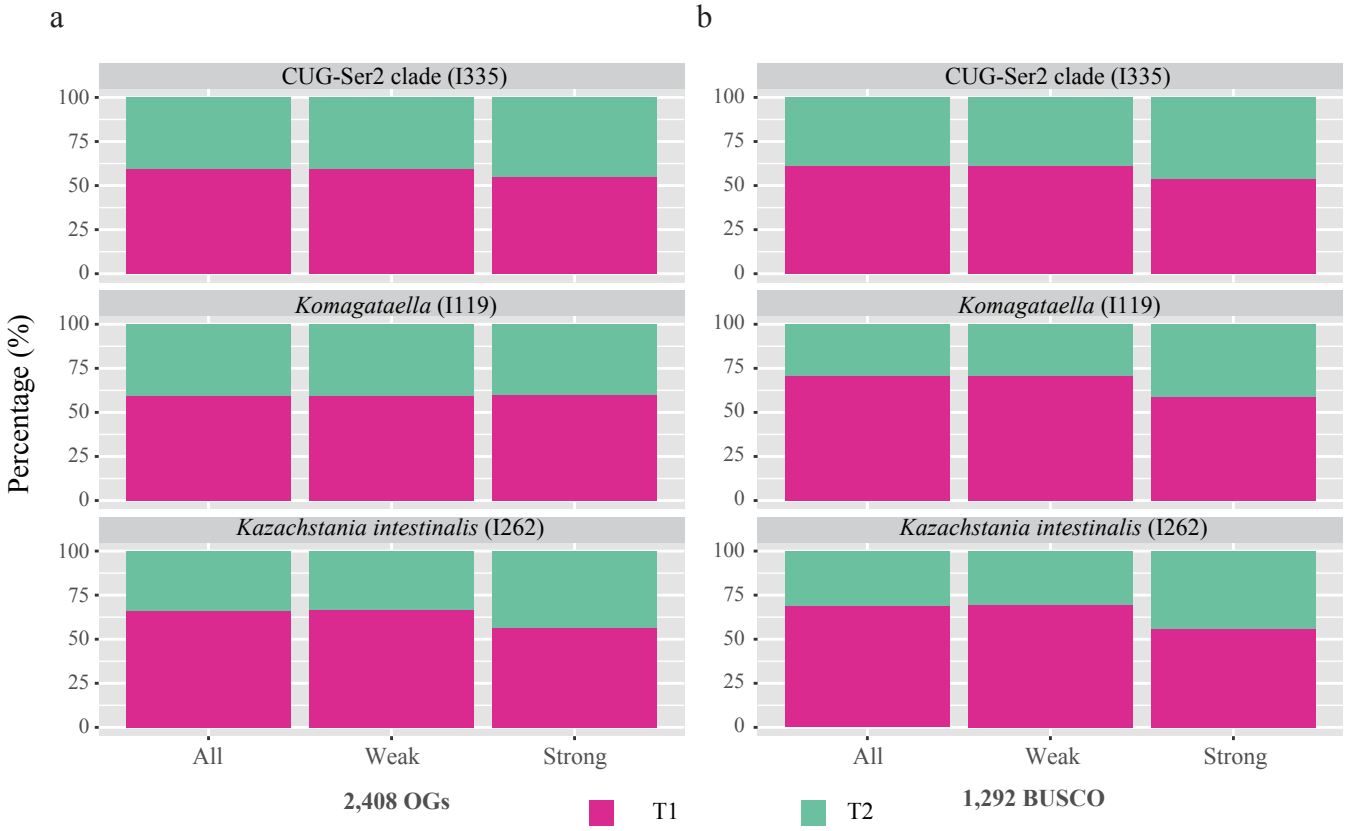
Additional Figure 21. Relationship between gene number, genome size, and number of gene duplicates among the 332 budding yeast genomes. The X axis shows gene number, the Y axis shows number of gene duplicates, and the Z axis genome size. The two newly sequenced genomes thought to correspond to hybrids are shown in red font; they have the highest numbers of genes, the highest numbers of gene duplicates, and large genome sizes.



Additional Figure 22. Distributions of gene-wise phylogenetic signal for 3 internal branches in the 2408OG data matrix (panel a) and the 1292BUSCO data matrix (panel b). For each branch for a given data matrix, the phylogenetic signal was calculated by measuring the difference in gene-wise log-likelihood scores (ΔGLS) for T1 versus T2. Pink bars denote genes supporting T1, whereas green bars denote genes supporting T2. The specific T1 and T2 topologies compared in each of the branches examined are provided in Additional Table 4 and shown in Additional Figure 19. The detailed values of the distributions of gene-wise phylogenetic signal in the 3 internal branches examined are provided in Additional Table 5.

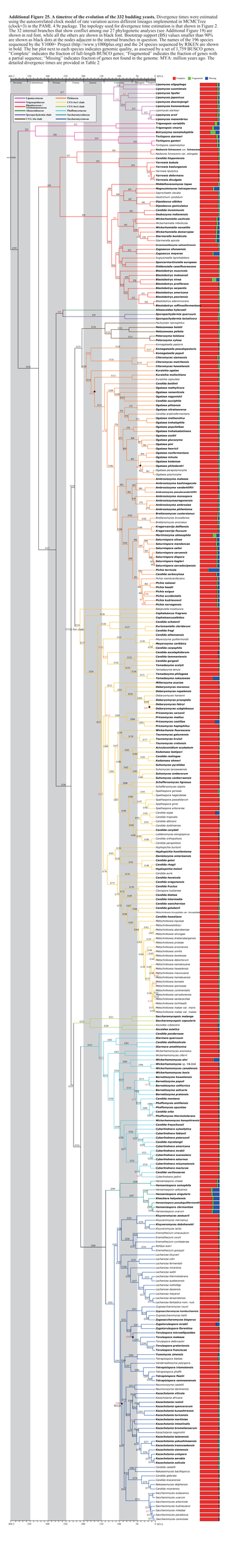


Additional Figure 23. Distributions of ranked gene-wise phylogenetic signal for 3 internal branches in the 2408OG data matrix (panel a) and the 1292BUSCO data matrix (panel b). For each branch for a given data matrix, the phylogenetic signal was calculated by measuring the difference in gene-wise log-likelihood scores (ΔGLS) for T1 versus T2. Pink bars denote genes supporting T1, whereas green bars denote genes supporting T2. The specific T1 and T2 topologies compared in each of the branches examined are provided in Additional Table 4 and shown in Additional Figure 19. The detailed values of the distributions of gene-wise phylogenetic signal in the 3 internal branches examined are provided in Additional Table 5.



Additional Figure 24. Percentage of sites supporting each of two alternative topological hypotheses for each of 3 internal branches in the 2408OG data matrix (panel a) and the 1292BUSCO data matrix (panel b). For each branch for a given data matrix, we evaluated all sites, weak sites with absolute ΔSLS (i.e., difference in site-wise log-likelihood scores for T1 against T2) smaller or equal to 0.5, and strong sites with absolute ΔSLS > 0.5. Pink bars denote genes supporting T1, whereas green bars denote genes supporting T2. The specific T1 and T2 topologies compared in each of the branches examined are provided in Additional Table 4 and shown in the Additional Figure 19.

Additional Figure 25. A timetree of the evolution of the 332 budding yeasts. Divergence times were estimated using the autocorrelated clock model of rate variation across different lineages implemented in MCMCTree (clock=3) in the PAML 4.9e package. The topology used for divergence time estimation is that shown in Figure 2. The 32 internal branches that show conflict among our 27 phylogenetic analyses (see Additional Figure 19) are shown in red font, while all the others are shown in black font. Bootstrap support (BS) values smaller than 90% are shown as black dots at the nodes adjacent to the internal branches in question. The names of the 196 species sequenced by the Y1000+ Project (http://www.y1000plus.org) and the 24 species sequenced by RIKEN are shown in bold. The bar plot next to each species indicates genome quality, as assessed by a set of 1,759 BUSCO genes. "Complete" indicates the fraction of full-length BUSCO genes; "Fragmented" indicates the fraction of genes with a partial sequence; "Missing" indicates fraction of genes not found in the genome. MYA: million years ago. The detailed divergence times are provided in Table 2.



Additional Figure 26. Visual comparison of Bayesian and RelTime divergence time estimates. The MCMCTree timetree is shown in blue and the RelTime timetree is shown in red. MYA: million years ago. Note that, consistent with previous work (Mello et al. 2017), the RelTime estimates were generally older than MCMCTree estimates for the deep internodes of the budding yeast phylogeny (e.g., for the internodes between the 12 major clades) and were generally younger than the MCMCTree estimates for shallower internodes (e.g., within the families Pichiaceae, Saccharomycodaceae, Saccharomycetaceae, Phaffomycetaceae, the CUG-Ala clade, and the CUG-Ser1 clade. Timetrees in Newick format are provided in the Figshare depository.