Supplementary Note S3

Estimation of relative timing of speciation events.

Phylogenetic trees drawn using ancestral loci at which single-copy syntenic orthologs have been retained in all post-WGD species (Class 4 in Fig. 2 in main text and Fig. S3.1 at right), can be used to determine the relative timing of post-WGD speciation events. Ancestral loci that have retained duplicates (Class 0 in Fig. 2 and Fig. S3.1 left) are not suitable for this purpose as they may undergo a period of relaxed selection following duplication[1,2], thus violating the assumptions of the molecular clock. They can be used however to estimate the time of divergence of duplicates created by WGD (at the common ancestor of the 'A' and 'B' copies; Fig. S3.1 at left).

This supplemental material describes a procedure to merge information from trees of duplicated and single-copy ancestral loci to produce a linear time-scale, on which 0 indicates the initial time of duplicate divergence and the timing of post-WGD speciation events are expressed as a proportions of the total time from duplicate divergence to *S. cerevisiae.*

1) Alignments of duplicated and single-copy syntenic ancestral loci

We used YGOB (http://wolfe.gen.tcd.ie/ygob/) to assemble sets of ancestral loci at which all post-WGD species had either retained two gene copies (Fig S3.1 at left), or had retained the same syntenic copy (Fig S3.1 at right). We discarded ribosomal proteins, ancestral loci at which one or more pre-WGD species possessed no ortholog and any ancestral loci for which no unambiguous *C. albicans* ortholog could be detected by reciprocal best blast hits with the *K. lactis* protein. The remaining 88 duplicated and 909 single-copy loci were aligned with ClustalW (default parameters), stripped of gapped columns and then merged to produce two super-alignments. The alignment of single-copy loci (referred to as A1 in this supplemental material) consists of 359,481 sites and the alignment of duplicated loci (A2) consists of 33,073 sites.

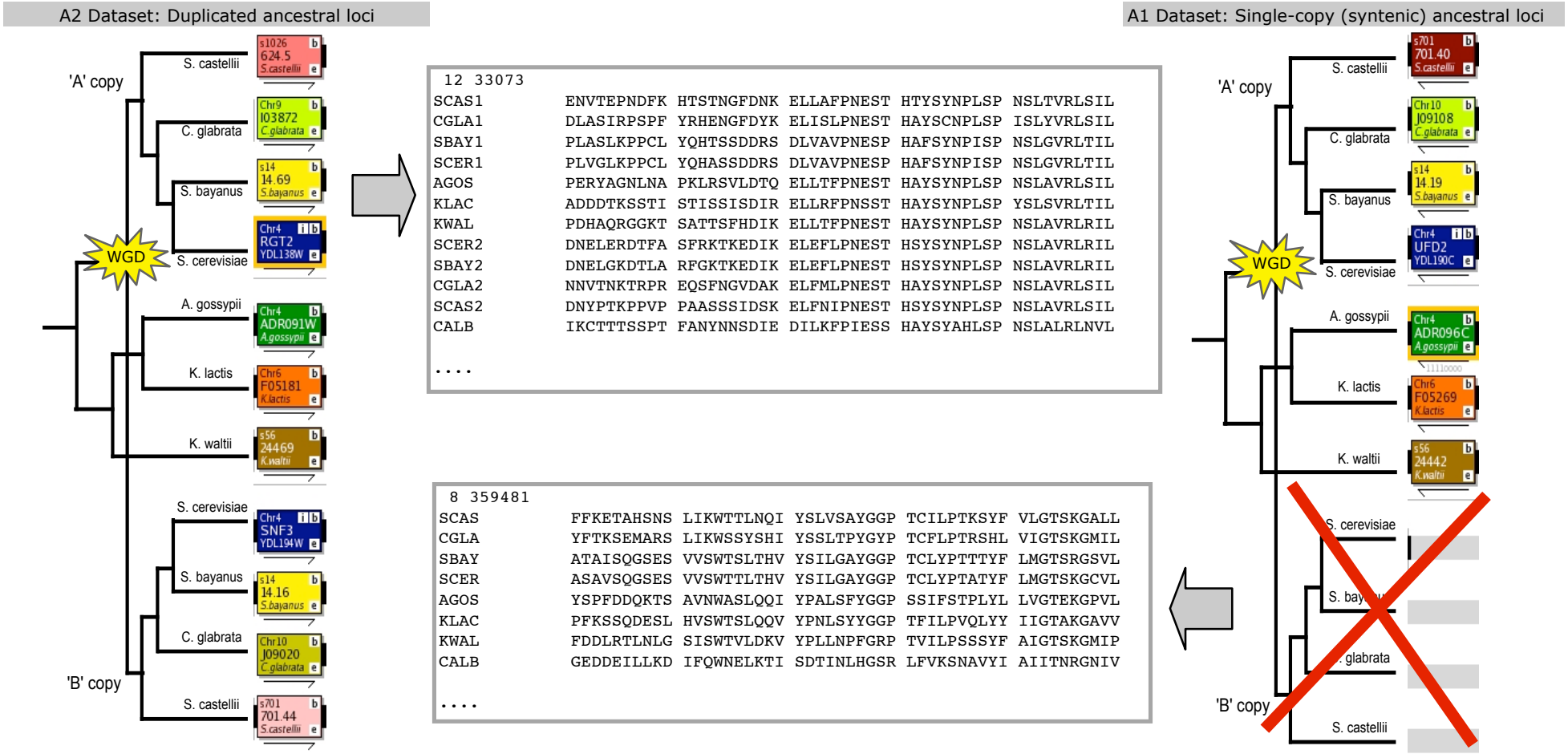


Figure S3.1 Assembly of alignments of ancestral loci that are still duplicated in all post-WGD species (left) and ancestral that have retained single-copy syntenic orthologs in all post-WGD species (right) using YGOB. The tree topologies on which these alignments are later evaluated are also shown.

2) Residue matching to construct comparable alignments of columns from duplicated and unduplicated ancestral loci

In order to merge information from trees drawn from duplicated and single-copy loci, we derived two new alignments (A1' and A2') by selecting pairs of columns from A1 and A2 that share the same amino acids in the pre-WGD taxa *K. waltii, K. lactis* and *A. gossypii* (Fig. S3.2). 71 columns of 33,073 in A2 (0.21%) could not be paired with columns in A1 and were excluded (Red columns in Fig. S3.2). A1' and A2' are therefore exactly the same length and consist of sites that (with the exception of duplication in some taxa) have similar evolutionary trajectories. Because A1' and A2' are large (33,002 sites) stochastic errors due to the residue-matching procedure should be negligible and the unduplicated regions of trees drawn from these alignments should be almost identical.

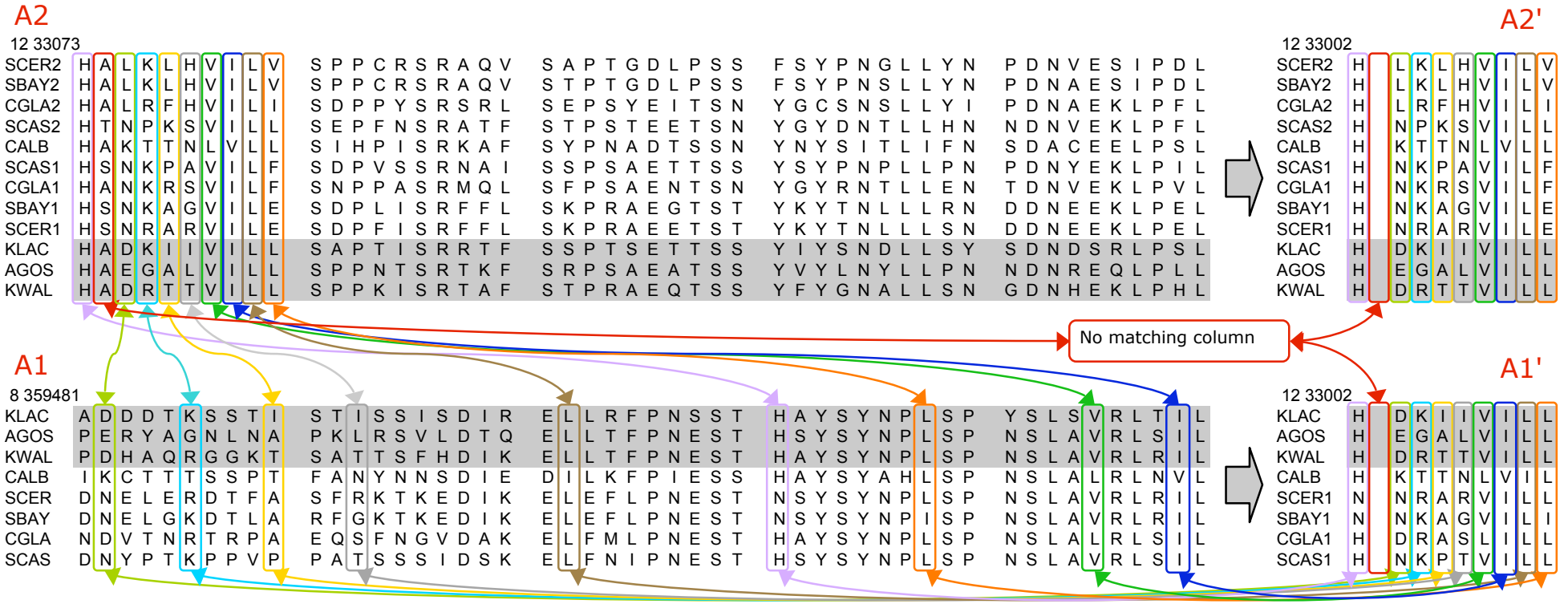
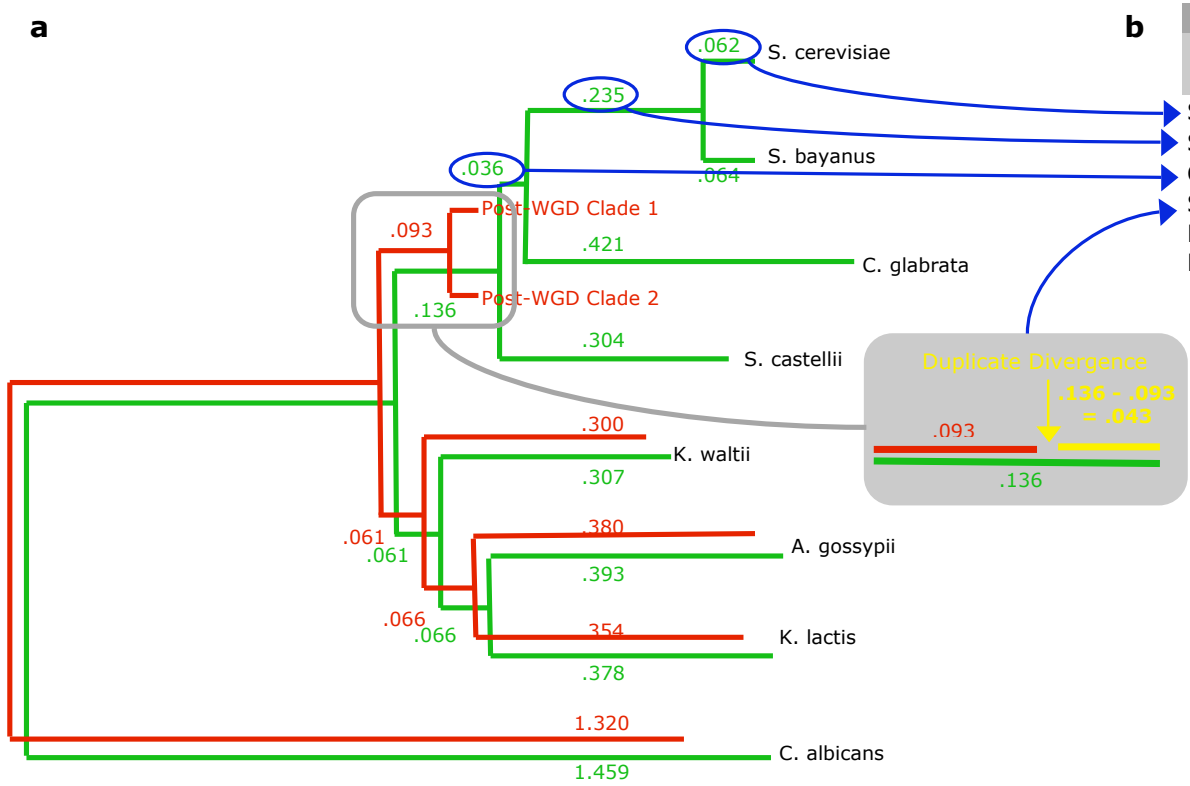


Figure S3.2 50 column example of the residue-matching procedure to construct comparable alignments (A1' and A2') of columns from ancestral loci that have been retained in duplicate in all post-WGD species and columns from ancestral loci that have retained single-copy syntenic orthologs in all post-WGD species. The taxa used for residue-matching are shaded in grey. The 10 boxed columns in A1 and A2 that are joined by arrows are examples of columns that have been "matched" between the two alignments. The column in A2 boxed in red could not be matched to a column in A1 (there is no 'AAA' in the 50 columns shown) and so has been omitted from the derived alignments A2' and A1'.

3) Timing of post-WGD speciation events since duplicate divergence

Maximum likelihood branch-length estimation was carried out for A1' (tree T1; green tree in Fig. S3.3) and A2' (tree T2; red tree in Fig. S3.3) under the topologies shown in Figure S3.1. As expected, unduplicated regions of T1 and T2 are very similar: In the pre-WGD clade T2 branches are on average 97.6% (range 93%-100%) of the length of the equivalent T1 branch, compared to 83.4% (range 79%-92%) for trees drawn from A1 and A2. The internal branches in the pre-WGD clade are exactly the same length.

Because the unduplicated regions of T1 and T2 are almost identical, we use the branch on T2 immediately prior to duplicate divergence to partition the branch on T1 between the divergence of the pre-WGD clade and the divergence of *S. castellii* into "pre-duplication" and "post-duplication" sections (Fig. S3.4 grey box). On this basis, the initial divergence of duplicates created by WGD occurred at a time equivalent to 4.3% amino acid divergence prior to the divergence of *S. castellii.* We use this figure, and the interspeciation branches on the post-WGD section of T1 (circled in blue), to estimate the relative timing of speciation events (Fig S3.3 b).



b

Time since duplicate divergence			
Divergence	Branch Length	Accumulated Branch Length	Proportion of Total Length
S. cerevisiae	0.062	0.376	1.00
S. bayanus	0.235	0.314	0.84
C. glabrata	0.036	0.079	0.21
S. castellii	0.043	0.043	0.11
Divergence of Duplicates	0	0	0.00

Figure S3.3 (a) Maximum likelihood trees, T1 and T2, drawn using A1' (green) and A2' (red; duplicated clades have been omitted for clarity). Model selection was performed using ProtTest and the model WAG + G + I + F was selected for all analyses. The gamma distribution was approximated with 8 rate classes. Trees were constrained to the topologies shown in Figure S3.1 and evaluated using Tree-Puzzle[3]. The topology of the post-WGD clade was determined as described in Supplementary Information 2. The topology and existence of the pre-WGD clade was inferred from additional trees drawn with A1 (data not shown). (b) Construction of a linear time-scale along the lineage from duplicate divergence to *S. cerevisiae* .

4) Confidence estimation for inferred speciation times

We calculated errors-bars for speciation time estimates by generating 100 bootstrap replicates of A2 and then performing the residue-matching procedure described above on each pseudo-replicate. Because there are 10 times more sites in A1 than A2, but only the number that can be paired are used, we are effectively also bootstrapping A1. The table below reports the mean and standard deiviation for each of the branches on the lineage from duplicate divergence to *S. cerevisiae.* In all cases, the standard deviation is small but greater than the difference between the mean and real data, suggesting that our estimates are likely to be robust.

Summary statistics for pseudo-replicates			
Divergence	% Time (Real Data)	% Time (Mean of Bootstraps)	Standard Deviation
S. cerevisiae	1.00	1.00	0.00
S. bayanus	0.84	0.84	0.01
C. glabrata	0.21	0.21	0.02
S. castellii	0.11	0.11	0.02
Divergence of Duplicates	0.00	0.00	0.00

References

1. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. Genome Biol 3, RESEARCH0008 (2002).

2. Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res 12, 1370 6 (2002).

3. Schmidt, H.A., K. Strimmer, M. Vingron, and A. von Haeseler (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502-504.